DOCUMENT RESUME

ED 268 158 TM 860 180

AUTHOR Marco, Gary L.; And Others
TITLE An Evaluation of Three Approximate Item Response Theory Models for Equating Test Scores.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-85-46
PUB DATE Dec 85
NOTE 87p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS College Entrance Examinations; Comparative Analysis; Computer Software; *Equated Scores; Estimation (Mathematics); High Schools; *Item Analysis; *Latent Trait Theory; *Mathematical Models; Scaling; Statistical Analysis; Statistical Studies
IDENTIFIERS Rasch Model; Scholastic Aptitude Test; *Three Parameter Model

ABSTRACT

Three item response models were evaluated for estimating item parameters and equating test scores. The models, which approximated the traditional three-parameter model, included: (1) the Rasch one-parameter model, operationalized in the BICAL computer program; (2) an approximate three-parameter logistic model based on coarse group data divided into fifths and twentieths, and using the Quantile modification of the LOGIST program; and (3) a modified three-parameter logistic model with fixed a's and c's, using the LOGIST computer program. The data came from a study of the Scholastic Aptitude Test (SAT), which involved the chain equating of a test to itself through five intermediary forms; approximately 2,670 cases were used for each SAT form. Results showed that item calibrations based on twentieths were closer to the true values and to LOGIST estimates than those based on fifths, but the equating results based on twentieths were not more accurate. Method (2) yielded highly accurate score conversions in equating a test to itself, and all three models yielded very accurate equating results. Questions were raised about the adequacy of equating a test to itself as a criterion for evaluating equating results. Further research was recommended before adopting any of the approximate models. Twelve tables and 22 figures are appended. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-85-48

ED268158

RESEARCH REPORT

AN EVALUATION OF THREE APPROXIMATE ITEM RESPONSE THEORY MODELS FOR EQUATING TEST SCORES

Gary L. Marco
Marilyn S. Wingersky
James B. Douglass

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Widenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Educational Testing Service
Princeton, New Jersey
December 1985

TM 860 180

An Evaluation of Three Approximate Item Response Theory Models for Equating Test Scores

Gary L. Marco
Educational Testing Service

Marilyn S. Wingersky
Educational Testing Service

James B. Douglass
Opinion Research Corporation

Educational Testing Service
Princeton, N.J.

December 1985

Copyright © 1985. Educational Testing Service. All rights reserved.

An Evaluation of Three Approximate Item Response Theory Models for Equating Test Scores

Abstract

The primary purpose of this study was to determine the extent to which three item response theory (IRT) models could be used to approximate the three-parameter logistic model in estimating item parameters and in equating test scores. These approximate models were less expensive to apply and in some cases used less data than the full-blown three-parameter model.

The approximations to the three-parameter model used in this study were (1) the Rasch one-parameter model, as operationalized in the BICAL computer program, (2) an approximate three-parameter logistic model based on grouped data divided into fifths and twentieths, and (3) a modified three-parameter logistic model with fixed a's and c's. The LOGIST computer program was used to estimate parameters for the modified three-parameter model; Quantile, a modified version of LOGIST that accepted coarsely grouped data, was used to estimate item parameters for the approximate three-parameter model.

In the case of the approximate models involving BICAL and LOGIST, results of separate item calibrations were used to place item parameter estimates on the same scale. In the case of the approximate model involving Quantile, a method of scaling the item parameter estimates indirectly through existing SAT scaled scores was used.

The data for the study came from a recent study (Petersen, Cook, & Stocking, 1983) of scale stability for the Scholastic Aptitude Test. As in the previous study, this study involved the chain equating of a test to itself through five intermediary forms. The sample consisted of approximately 2,670 cases for each of the SAT forms used.

The results of the study were as follows: (1) the item calibrations based on twentieths were closer to the true values and to LOGIST estimates than item calibrations based on fifths; (2) the equating results based on twentieths, however, were not more accurate generally than those based on fifths; (3) the three-parameter model using coarse groupings yielded highly accurate score conversions in equating a test to itself, more accurate in fact than the full-blown three-parameter models studied by Petersen, Cook, and Stocking; and (4) all of the approximate models yielded very accurate equating results. A follow-up analysis indicated that these unexpected equating results were due in large part to the indirect method used to place item parameter estimates on scale through existing score conversions derived from conventional equating methods. The success of the approximate models raises a question about the adequacy of equating a test to itself as a criterion for evaluating equating results. Further research is recommended before any of the approximate models are used operationally.

An Evaluation of Three Approximate Item Response Theory Models for Equating Test Scores[1]

The increasing internal and external demands made on testing programs have underscored the inflexibility of score equating methods used traditionally. Item response theory (IRT) equating offers several advantages in this context, including improved equating (particularly at the ends of the scale), greater test security through less dependence on items common to a particular previously used form, and easier re-equating when items are added or deleted. While these are important advantages, test disclosure legislation has created a more urgent need for IRT-based equating. The New York State test disclosure legislation requires that those items on which reported scores are based be made available to the public. An important advantage of using IRT methods in response to such legislation is that equating based on item pretest data is possible prior to a test's administration (pre-equating), thus permitting forms to be developed without requiring a special equating administration.

Although IRT methods of equating are superior to traditional methods in a number of important respects (see Marco, Petersen, & Stewart, 1983), the costs of converting from traditional to IRT equating methods can be substantial. The LOGIST computer program (Wingersky, 1983) and other computer programs used to estimate IRT item parameters for the three-parameter logistic test model take a considerable amount of computer time and thus are expensive to run for large data sets. The costs are particularly high when IRT methods are introduced into an existing testing program because the parameters of a large number of items must not only be estimated using a program like LOGIST but also placed on a common scale through complicated common-item linkages.

Previous research suggests that approximate IRT methods might be useful when the objective is to equate test scores. In a study of PSAT/NMSQT pre-equating, Marco (1977) used an approximate method for placing item parameter estimates on a common scale to avoid the considerable expense of calibrating items from a large number of test forms. He used existing score equating results based on traditional linear equating to scale item parameter estimates from separate applications of LOGIST. Marco found that, except at the upper end of the score scale, the pre-equating results agreed reasonably well with the criterion equatings. Different item calibration techniques have also been compared. In a simulation study Ree (1979) compared item parameter and ability estimates obtained from LOGIST and two of Urry's programs, ANCILLES and OGIVIA, with the parameters from which the simulated data had been generated. He found none of the programs uniformly superior for parameter estimation. However, the cost of using Urry's programs were only 10% to 15% of the cost of using LOGIST.

Several studies have evaluated the Rasch model, the simplest IRT model, for equating test scores. Rentz and Bashaw (1975) equated scores on a number of elementary school reading tests with the Rasch model. They found good agreement between the equating results of the equipercentile and Rasch models. In another study Douglass (1980) found that the Rasch model provided better equating results than the two-parameter logistic model for a classroom achievement testing system. The Rasch equatings were more consistent across different-sized examinee samples. There was also evidence that, compared to two-parameter equatings, Rasch equatings tended to result in less equating error when dissimilar examinee samples were used. In a large scale study of score equating methods, Marco, Petersen,

and Stewart (1983) compared one-parameter and three-parameter logistic, equipercentile, and linear score equating models under varying conditions. They found that when a test was equated to itself using random samples, all of the equating models had a small amount of equating error. But when dissimilar samples were used, both the one- and three-parameter logistic models were clearly superior. However, when a test was equated to a test differing in difficulty, the equating results of the one-parameter model were unsatisfactory. In another study of score equating models, Kolen (1981) compared linear, equipercentile, and one-, two-, and three-parameter logistic score equating models. Like Marco, Petersen, and Stewart, he found that the one-parameter logistic model yielded inadequate results for equating tests of unequal difficulty. Other studies (e.g., Slinde & Linn, 1978; Loyd & Hoover, 1980; and Holmes, 1982) have also evaluated the adequacy of the Rasch model for score equating, with mixed results.

These studies from the IRT research literature support the possible utility of using approximate methods, but also call attention to conditions under which approximate methods might give unsatisfactory results. For a test that has little form-to-form variation and only moderate differences in the ability of the examinees from one administration to another, there is good reason to expect that approximate methods might provide acceptable results at a much lower cost. Of course, approximate methods would be most useful in small testing programs, which cannot afford to use the more expensive methods.

The primary purpose of this study was to determine the extent to which the many advantages of IRT score equating could be realized by using approximate models that were less expensive to apply and, in some cases, required less data than the full-blown three-parameter logistic model, as operationalized in the LOGIST computer program. Three IRT equating models intended to approximate the three-parameter logistic equating model were studied. Various

hypotheses concerning these models were formulated. These hypotheses are outlined in the first part of the section on results.

Procedures

The tests and examinee samples for this study were those used for a recent Scholastic Aptitude Test (SAT) scale stability study (Petersen, Cook, & Stocking, 1983). That study investigated several methods for equating scores from six SAT-verbal and six SAT-mathematical test forms. Included among the methods were linear equating, equipercentile equating, and several variants of IRT equating. The study involved the chain equating of a test to itself through five intermediary forms. The current study builds on these results by providing data on a number of additional equating methods intended to approximate three-parameter logistic equating.

Tests and Test Scores

The tests consisted of six operational and six equating tests for SAT-verbal and SAT-mathematical, respectively, administered between December 1973 and May 1979. The tests were chosen so that the equatings formed a closed circle in that a test form could be equated to itself. These tests are identified in Figure 1, which shows the chains of six verbal and six mathematical equatings that were used in the study. SAT forms are indicated by upper case letters and equating tests by lower case letters. Each SAT-verbal form except Form V4, which was administered prior to the Fall of 1974, had 85 items; Form V4 contained 90 items. A given verbal equating test contained 40 items. Each SAT-mathematical form except Form Y3 had 60 items; and each mathematical equating test except

Form fn, 25 items. Due to printing error, SAT-mathematical Form Y3 had 59 items, and mathematical equating test fn had 24 items.

- - - - - - - - - - - - -

Insert Figure 1 about here

- - - - - - - - - - - - -

The SAT was shortened from 75 minutes to 60 minutes in the fall of 1974 to permit the administration of the SAT's companion test, the Test of Standard Written English. The shorter SAT-verbal forms contained the same item types as the previous forms, but the numbers of items (all five-choice) within a given item type were changed. The shorter SAT-mathematical form contained quantitative (four-choice) comparisons and regular mathematics (five-choice) items instead of data sufficiency (five-choice) and regular mathematics (five-choice) items.

Raw scores on the SAT are formula scores based on the number right minus a fraction of the number wrong, where the fraction is 1/(no. of response options - 1). Raw scores for a particular test form are converted to scaled scores on the 200 to 800 College Board scale by applying the mathematical transformation derived through score equating.

Data Used in the Study

The sample consisted of approximately 2,670 cases for each pairing of an SAT form and an equating test shown in Figure 1. The actual sample sizes ranged from 2,527 to 2,879. The samples were randomly selected from examinees taking the SAT at the respective administrations. Figure 2 shows the data sets that were used in the study. Individual records contained item response data appropriate for use in the various computer programs, which required information on right and wrong responses to each test item. Records also contained information

on items omitted and items not reached. Table 1 gives the SAT scaled score means and standard deviations for the samples used in the study.

- -

Insert Figure 2 and Table 1 about here

- -

Equating Design

Figure 1 shows the chains of six verbal and six mathematical equatings that were used in the study. These chains were also used in the SAT scale stability study. In that study and in this one SAT-verbal Form V4 was equated to itself through several intermediary forms. Form V4 was treated as the base form of the test for equating scores on Form Z5 to scores on Form V4. Form Z5 in turn was treated as the base form for equating scores on Form Y2 to scores on Form Z5. In the last step scores on Form V4 were equated to scores on Form X2 using Form X2 as the base. The results of this chain equating could be compared to the original scores. Ideally, the results would be identical. Any discrepancy could be attributed to the particular equating method used. In the 1981 study all equatings made use of common item linkages established by the equating tests. In the current study some equatings depended upon the equating test data, and some did not.

The idea of equating a test to itself as a way of evaluating equating methods was introduced by Levine (1955) when he developed several linear true score equating methods. Marco, Petersen, and Stewart (1983), in their study of curvilinear equating methods, also used this type of criterion. This idea was extended by Petersen, Cook, and Stocking (1983) to chain equating, whereby a test is equated to itself through a series of intermediate forms. In this way

variations in test length, test difficulty, etc., can be introduced to discover to what extent various equating models can adapt to changing conditions.

The reason that equating a test to itself is such a powerful idea is that when a test is equated to another test, the true relationship of the scores is not known. Thus, when studying equating in a natural setting, equating a test to itself is the only way to ensure the availability of a known criterion. When a test is equated to a different test, simulations can be used to establish a known criterion, but then it is difficult to introduce the kind of variation that exists naturally.

In the equating chain used for this study and the SAT scale stability study, SAT forms differed systematically only in that Form V4 was administered before the time limit for SAT-verbal or -mathematical was changed from 75 minutes to 60 minutes. This decrease in time limits necessitated a change in the mixture of item types in SAT-verbal and the introduction of Quantitative Comparison items in SAT-mathematical. These changes plus the natural variation from form to form probably introduced some curvilinearity into the equating relationship, and some differences in reliability could be expected from the changes in test lengths.

Equating Models

There are three separate but related steps required to use item response theory for score equating. The first step, item calibration, is to estimate item parameters. The second step, item parameter transformation - required when item parameters are estimated in separate computer runs, is to place item parameters on a common scale. The third step, score equating, is to relate raw scores on various pairs of tests to underlying abilities. In IRT true score

equating, the only equating method used in this study, scores on two tests are considered to be equated if and only if the true scores correspond to the same underlying ability level. However, various item calibration and item parameter transformation procedures were used. The type of data set (data from two SAT forms and an equating test or data from one SAT form) also varied, depending on the equating method.

The approximations to the three-parameter logistic model used in this study were (1) the Rasch one-parameter logistic model, (2) an approximate three-parameter logistic model based on groups divided into fifths and twentieths, and (3) a modified three-parameter logistic model with fixed $\underline{a}$'s and $\underline{c}$'s. The one-parameter model was included in the study for comparative purposes because of its relative simplicity and its wide use in some professional circles. The BICAL computer program was used to estimate item parameters for the one-parameter model; and LOGIST, for the modified three-parameter model. Quantile, a modified version of LOGIST written for this study, was used to estimate item parameters for the approximate three-parameter model.

If item parameters are estimated in separate computer runs, the scales underlying the estimates will, according to the theory, differ by a linear transformation. Thus, before such estimates can be used for score equating, they must be transformed to a common scale. This can be accomplished in several ways. In this study item parameters were placed on a common scale (1) by calibrating concurrently items from the pairs of test forms whose scores were to be equated and their common equating tests; (2) by equating the $\underline{b}$'s, the item difficulties, using parameter estimates for the equating test items from separate item calibrations; and by equating $\underline{\theta}$'s, examinee abilities, indirectly using existing operational score equating parameters. The third procedure had been used by Marco (1977) to equate item parameter estimates from different samples when there is no equating test.

Table 2 shows the variations associated with these three approximate equating models. The samples sizes were the same for the three models - approximately 2,670 cases for each sample. Also, the same type of equating was used in each case; namely, IRT true score equating. This kind of equating is described at the end of this section. The approximate equating models varied as to the type of data used, the method of item calibration, and the method of item parameter equating. The various data sets used in the study have already been identified in Figure 2.

- - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - -

One-parameter logistic (Rasch) model. The computer program BICAL (Wright & Mead, Note 2) was used to calibrate the items from the 12 verbal and 12 mathematical data sets shown in Figure 2. A separate application of BICAL was made for each data set. Since BICAL provides item parameters on a separate scale for each item calibration, the item parameter estimates had to be transformed. This was accomplished by setting equal the means from the two calibrations of the common items (Wright & Stone, 1979). (For this method an additive constant provides the appropriate adjustment to the b's.) For example, the item difficulty parameter estimates for SAT-verbal Form X2 and Equating Test fe were equated to the scale for SAT-verbal Form V4 and Equating Test fe by subtracting .090. This constant was found by subtracting the mean b for the items in the equating test fe for the examinees who took Form X2 from the mean b for the same items for the examinees who took Form V4. Table 3 gives the equating transformations used to place BICAL estimates on the Form V4 scale. The transformation

for placing the item parameter estimates for any particular form on the V4 scale was obtained simply by summing the constants in the chain. In this way, a transformation was obtained for equating the item parameters for Form V4 to the V4 scale indirectly. Ideally the constant would equal 0. This final transformation is given in the last line of the table.

- - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - -

Approximate three-parameter logistic model. This model was intended to approximate the three-parameter logistic model using grouped data. Previously, Bock (1976) had used coarse grouping in the computer program LOGOG to estimate item parameters and had obtained relatively accurate, albeit inconsistent, estimates with large samples. Considerable cost savings could result if the abilities for each examinee did not have to be estimated. This model was designed to calibrate items on the basis of item analysis information routinely produced at Educational Testing Service.

A new computer program Quantile was developed by modifying LOGIST to accept grouped data. The Quantile version estimates the item parameters for a test using item responses of groups of examinees instead of responses of individual examinees. The examinees are divided into groups before the program is applied. For each item the input to the program is the number of examinees in each group who answered the item correctly plus a fraction (1/(no. of response options)) of the number who omitted the item, the number of examinees in each group who reached the item, the total number of examinees who answered the item correctly, and the total number of examinees who omitted the item. All of the examinees in a group are treated as having the same ability. This ability is estimated using maximum likelihood in the same manner as individual abilities

are estimated by LOGIST. The options available and the output are identical to the LOGIST options and output.

All of the required information can be derived from routine ETS item analysis data from groupings based on fifths. For purposes of the study the special program Anytiles was written to allow groupings into both fifths and twentieths. The latter grouping was used for comparison purposes even though not produced by routine item analysis.

Quantile produces estimates of all three item parameters for the logistic test model. Before the program was used in the study, it was tested on artificial data. These test runs indicated that the a's and c's were underestimated when compared with their true values. When the c's were fixed at their true values, however, the a's were unbiased. (Previous comparisons of LOGIST results with true values from artificial data had demonstrated that LOGIST item parameter estimates based on individual data are unbiased when based on large samples.)

To correct for this bias, an empirical correction was computed using the item calibrations for the SAT-verbal and SAT-mathematical data from the March, May, and June 1982 administrations for which LOGIST item calibrations existed. Separate corrections for coarse groupings (fifths and twentieths) were computed for verbal items (five-choice), four-choice mathematical items, and five-choice mathematical items. These empirical corrections were derived in the following way: (1) The b parameter estimates from Quantile were equated to the b parameters from LOGIST to place them on the same scale. (Means and standard deviations were set equal.) (2) The a parameter estimates were equated to the a's from LOGIST by setting means and standard deviations equal after removing any pairs where either a was greater than 1.5. (3) Step 2 was repeated for the c's, removing any items with either c at the common value or

c greater than .4. For the c's that were not estimated but set equal to the common c, the mean c was compared to the mean c from LOGIST for these same items and a constant adjustment obtained. (4) Results for the March, May, and June data sets were compared and average linear transformations obtained. Table 4 gives the average corrections applied in the study.

- - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - -

The Quantile program was applied to item response data from the SAT forms; equating test data were not used. SAT-verbal data from the following data sets shown in Figure 2 were used: V4 and fe, X2 and fm, Y3 and fw, B3 and fk, Y2 and fu, Z5 and et, and V4 and et. The SAT-mathematical data sets consisted of data from V4 and ff, X2 and fn, Y3 and fu, B3 and fl, Y2 and fv, Z5 and eu, and V4 and eu.

Once these item calibrations were available, they were adjusted by applying the appropriate empirical corrections (see Table 4). The parameter estimates, both corrected and uncorrected, then had to be transformed to a common scale. This was accomplished using the operational score equating parameters in the manner described by Marco (1977). The essential steps were as follows:

1. For a given ability level, $\underline{\theta}_i$, for test i, compute the true number-right score, $\underline{R}_i$, by $\Sigma \underline{P}_g$, where $\underline{P}_g$, the probability of answering item g on test i correctly, is computed from the estimated parameters for item g.
2. Express $\underline{R}_i$ as a true formula score, $\underline{FS}_i$, under the assumption that all items are answered: $\underline{FS} = \underline{R} - (\underline{N}_5 - \underline{R}_5)/4$ for SAT-verbal (five-choice

items) and $\underline{FS} = \underline{R} - (\underline{N}_4 - \underline{R}_4)/3 - (\underline{N}_5 - \underline{R}_5)/4$ for SAT-mathematical (four- and five-choice items), where $\underline{N}$ is the number of test items and the subscript indicates whether four-choice (4) or five-choice (5) items are involved.

3. Transform $\underline{FS}_i$ to the College Board scale ($\underline{S}$) using the operational scaling parameters derived previously when the tests were originally equated (see Table 5): $\underline{S} = \underline{A}_i\underline{FS}_i + \underline{B}_i$.

4. Find the true $\underline{FS}$ on test j corresponding to this scaled score:

$$\underline{FS}_j = (\underline{S} - \underline{B}_j)/\underline{A}_j \quad ,$$

where $\underline{A}_j$ and $\underline{B}_j$ are the scaling parameters (see Table 5).

5. Compute the true $\underline{R}_j$ from the formulas in Step 2.

6. Determine the corresponding ability level $\underline{\theta}_j$ for test j by finding the $\underline{\theta}_j$ for which $\Sigma\underline{P}_h$ equals $\underline{R}_j$ using the item parameter estimates for test j.

7. Apply steps 1-6 to approximately 60 equally spaced ability levels between -3 and 3.

8. Determine the straight line relating the $\underline{\theta}_j$'s to the $\underline{\theta}_i$'s in the range -1.75 to 1.75 by setting their means and standard deviations equal. The range is restricted to prevent outliers in the tails of the score range from influencing the results. This process results in a transformation $\underline{\theta}_i = \underline{A}\,\underline{\theta}_j + \underline{B}$. Here $\underline{A} = SD(\underline{\theta}_i)/SD(\underline{\theta}_j)$ and $\underline{B} = M(\theta_i) - \underline{A}M(\theta_j)$, where M and SD stand for mean and standard deviation, respectively.

9. Determine the transformation for placing the item parameter estimates for test j onto the scale of the item parameter estimates for test i as follows:

$$\underline{b}_i = \underline{A}\underline{b}_j + \underline{B} \text{ and } \underline{a}_i = \underline{a}_j/\underline{A} \ .$$

The $\underline{c}$'s are unaffected by scale transformations.

- - - - - - - - - - - - - -

Insert Table 5 about here

- - - - - - - - - - - - - -

Table 6 gives the final transformations to the Form V4 scale for item difficulties estimated from the Quantile computer program. Corrected transformations were not determined for SAT-mathematical, for the SAT-verbal results (see the section on results) indicated that the corrections did not improve the equating transformations. Ideally, the transformation from Form V4 to the Form V4 scale would be $(1.0 \times \underline{b}) + 0$; that is, the original value would be returned.

Once the item parameter estimates were transformed to the V4 scale, the scores on Form V4 could be equated to themselves. For this purpose only the two sets of V4 item parameters estimated were utilized; the use of intermediate estimates were unnecessary.

- - - - - - - - - - - - - -

Insert Table 6 about here

- - - - - - - - - - - - - -

Modified three-parameter logistic model. A reasonable alternative to the three-parameter logistic model is a modified version that involves fixing the item discriminations at a common value and fixing the lower asymptotes at a common non-zero value for all items. This model has two potential advantages over the one-parameter logistic model discussed previously. First, for multiple-choice items a lower asymptote greater than zero is a more reasonable assumption

than a lower asymptote of zero. Second, the LOGIST program takes omitted and not reached items into account, whereas the BICAL program considers omitted and not reached items as incorrect responses.

LOGIST was used to estimate item parameters for the modified three-parameter model. Item parameters a and c were fixed at .788 and .149, respectively, for SAT-verbal and at .898 and .113 (for five-choice items) or .155 (for four-choice items), respectively, for SAT-mathematical. The fixed values were averages from previous SAT item calibrations from LOGIST.

Each item calibration involved two samples of examinees, one which took an SAT and an equating test and one which took a different SAT form but the same equating test. For example, data from examinees who took SAT Form V4 and Equating Test fe were merged with the data from examinees who took SAT Form X2 and Equating Test fe for calibration purposes (see Figure 2). The LOGIST computer program permits this type of calibration, even though all examinees do not answer all items, and returns item parameter estimates on the same scale for both SAT forms and the equating test. Thus, there is no need to derive a separate transformation to equate the item parameters from the two forms.

Score equating was accomplished by equating successively the scores of the forms represented in the concurrent item calibrations. Thus, transforming item parameter estimates to a common scale was unnecessary. The scores for Form X2 were equated to the scores for Form V4 using the item calibration involving these two forms, and raw-to-scaled-score conversions were obtained. Then Form X2 scores transformed to the V4 scale were used as input for the equating of scores on Form Y3 to scores on Form X2. This process was continued until the chain ended with the equating of scores on Form V4 to scores on Form Z5. This final raw-to-scaled-score conversions for Form V4 could then be compared with the original conversions.

Score Equating Method: Curvilinear True Score Equating

Once appropriate item parameter estimates were available for the approximate methods, score equating could be accomplished. IRT true score equating was used in each instance. This method was also used in the SAT scale stability study. Lord (1980) discussed this kind of equating in his recent book. Here only a brief summary of the method is given.

The first step in the score equating process is to compute the true number-right score from the item parameters that have been placed on a common scale. The true number-right score is a function of the ability level θ and the item parameters a, b, and c. For the one-parameter logistic model $a = 1/1.702$ and $c = 0$ for every item. (The division by 1.702 is necessary to make the results of BICAL consistent with the results of LOGIST). Let R stand for the number right true score. Then $R(\theta) = P_g(\theta)$, where

$$P_g(\theta) = c_g + (1 - c_g)/(1 + \exp(1.702\, a_g\, (\theta - b_g))).$$

The true formula score is obtained by assuming that everyone answered all items, so that

$$FS = R - (N - R)/(k_g - 1),$$

where N is the number of test items and k_g is the number of response options for the item g. In the case of SAT-mathematical, which contained both four- and five-choice items two correction terms were used - one for four-choice items and one for five-choice items.

True formula scores on two tests are said to be equated if they are functions of the same θ. To obtain the raw-to-scaled score conversions for raw scores, it is necessary only to determine θ for each FS on the test form

and to find the corresponding FS on the other form. In practice, of course, estimates of the item parameters rather than the unknown true parameters are used for calculating the true scores.

For the one-parameter logistic model and the approximate three-parameter logistic model, equating Form V4 to itself was accomplished in one step once the item parameters were transformed. For the modified three-parameter model, the results of the item calibrations were applied stepwise, starting with the equating of raw scores on Form V4 to raw scores on Form X2 (see Figure 1), and continuing around the circle by feeding in score conversions from the previous equating. At the end of the chain, initial raw scores on Form V4 were transformed to scaled scores by applying the original scaling parameters for Form V4 (see Table 5) to the equated raw scores corresponding to these initial raw scores. These "final" scaled scores could then be compared to the initial scaled scores.

Results and Discussion

A number of specific hypotheses were formulated for the study and expected to be confirmed: (a) The item parameter estimates from the groups based on twentieths more closely match the three-parameter logistic estimates from LOGIST than the estimates based on fifths. (b) The item parameter estimates corrected for coarse grouping more closely match the three-parameter logistic estimates from LOGIST than the uncorrected estimates. (c) The approximate three-parameter logistic equating model using item parameter estimates from the groupings based on twentieths yields less equating error than the model using estimates based on fifths. (d) The approximate three-parameter logistic equating model using item

parameter estimates corrected for coarse grouping yields less equating error than the model using uncorrected estimates. (e) The approximate equating models yield more equating error than the concurrent equating model. (f) The more complex approximate model - the modified three-parameter model - yields less equating error than the other approximate models. (g) Because it utilizes only item difficulty parameters, the one-parameter logistic model (Rasch) yields more equating error than any of the other approximate models.

In this set of hypotheses reference was made to the concurrent equating model, which was evaluated in the SAT scale stability study (Petersen, Cook, & Stocking, 1983), and, beginning in January 1982, is being used operationally to equate SAT scores. In concurrent IRT equating item parameter transformation is unnecessary as a separate step. Items from a new form, an old form, and a common anchor test are calibrated together using LOGIST, which produces item parameter estimates on a common scale. Then new form scores are equated to old form scores on the basis of these item parameter estimates. For the next equating the new form becomes the old form. Items from this form, another new form, and another common anchor test are calibrated together, and the scores on the total tests are equated. This kind of sequential "pairwise" equating was judged the most adequate of the three-parameter logistic IRT equating models represented in the SAT scale stability study. The results for concurrent equating referenced in this report are taken from that study.

Comparisons of Item Parameter Estimates

The item parameter estimates from the Quantile computer program were evaluated for a set of artificial data on 45 items and 1,500 examinees for which the true parameter values are known and LOGIST results already exist. The

Quantile item parameter estimates were compared to the true values and to the LOGIST estimates. Since for those data LOGIST parameter estimates are known to have negligible bias, the LOGIST results can be used as a criterion for determining bias in the Quantile estimates where the true item parameters are unknown. The abilities for each set of parameter estimates were standardized to a mean of 0 and standard deviation of 1 for abilities between -3 and 3, and the item parameters adjusted accordingly to put all of the parameters on a common metric.

The plots comparing the Quantile data to the true values are shown in Figure 3 for the fifths and Figure 4 for the twentieths. The circles on the a and c plots indicate items for which c was set to a common c value by the computer programs. The c's and the larger a's are underestimated. This same bias is evident in the plots in Figures 5 and 6, comparing the Quantile estimates to the LOGIST estimates for the fifths and the twentieths, respectively.

- -

Insert Figures 4, 5, and 6 about here

- -

Table 7 gives the summary statistics for the comparisons for the artificial data. The "mean absolute differences between the item response functions" is the absolute difference between the two curves averaged over all of the examinees and over all items. This is highest for the LOGIST results compared to true values. Asympototically, LOGIST minimizes the weighted mean squared error, not the mean absolute error. For all of the other statistics the LOGIST results agree better with the true values than do the Quantile results. The hypothesis that the twentieths give better estimates than the fifths is confirmed. The item-ability regressions in Figure 7 show the effects of using grouped data.

For Item 21 and most other items, the estimated curve fit the true curve very well. For Item 28 and a few other items, however, poor fit resulted.

- -

Insert Table 7 and Figure 7 about here

- -

For the two sets of data for Form V4 scale, Quantile results were compared to the LOGIST results using the uncorrected item parameter estimates and the estimates corrected for bias. The summary statistics are given in Table 8. Figures 8 to 11 show the comparisons for SAT Form V4 and equating Test fe for uncorrected and corrected item parameter estimates, respectively. Figures 12 to 15 show the comparisons for SAT Form V4 and Equating Test et. The hypothesis that the item parameter estimates corrected for coarse grouping more closely match the LOGIST results than the uncorrected estimates was confirmed only for the b's and c's of Form V4-et. For the a's and c's of SAT Form V4 and Equating Test f3, the uncorrected estimates agree better with the LOGIST results than do the corrected estimates, and for the b's the corrections had negligible effect.

- -

Insert Table 8 and Figures 8-15 about here

- -

Comparisons of Equating Results - Descriptions of the Tables and Figures

The equating results from the various approximate equating models are given in Table 9 and Figures 16 and 17 for SAT-verbal and in Table 10 and Figure 18 for SAT-mathematical. Tables 9 and 10 give the point-by-point conversions. In these tables the initial scaled score (the criterion) is the scaled score that

was obtained from the original equating. The final scaled score is the scaled score that was obtained by applying the original raw-to-scaled score conversion parameters for Form V4 to the equated raw scores resulting from the chain equating. The results are directly comparable with those obtained in the SAT scale stability study.

- -

Insert Tables 9 and 10 and Figures 16, 17, and 18 about here

- -

Figures 16, 17, and 18 present the equating results graphically. These graphs show the differences between the model results (final scaled scores) and the criterion (initial scaled scores). The codes used in the figures are as follows:

CRIT: Criterion--initial scaled score,

CONCUR: Concurrent (three-parameter logistic),

BICAL: BICAL (one-parameter logistic),

WO5: Approximate three-parameter logistic based on fifths without corrections to item parameter estimates,

W5: Approximate three-parameter logistic based on fifths with corrections to item parameter estimates,

WO20: Approximate three-parameter logistic based on twentieths without corrections to item parameter estimates,

W20: Approximate three-parameter logistic based on twentieths with corrections to item parameter estimates, and

MOD3: Modified three-parameter logistic.

Table 11 summarizes the point-by-point results in terms of several discrepancy indices. It gives the means and standard deviations for the scaled scores resulting from the various equatings and for the criterion scores. Ideally, the mean and standard deviation for a particular model would correspond exactly to the mean and standard deviation of the criterion scores.

- - - - - - - - - - - - - - -

Insert Table 11 about here

- - - - - - - - - - - - - - -

The table also gives the weighted mean squared difference and its two components, the mean difference and the standard deviation of the difference. For each raw score x on Form V4 there are final scaled scores resulting from the various equatings and an initial scaled score. The smaller the differences between the final score, $\underline{t}'_j$, for a particular equating model and the initial score, $\underline{t}_j$, the more accurate the equating model is. To compute the weighted mean squared differences the values of $\underline{x}$ are weighted according to their actual occurrence in some reference group. The weighted mean squared difference is equal to the variance of the difference plus the mean difference squared; that is

$$\Sigma \underline{f}_j \underline{d}_j^2 / \underline{n} = \Sigma \underline{f}_j (\underline{d}_j - \bar{\underline{d}})^2 / \underline{n} + \bar{\underline{d}}^2,$$

where $\underline{d}_j = (\underline{t}'_j - \underline{t}_j)$, $\underline{t}'_j$ is the estimated scale score for raw score $\underline{x}_j$, $\underline{t}_j$ is the initial or criterion scale score for x_j, $\underline{f}_j$ is the frequency of x_j, $\underline{n} = \Sigma f_j$, $\bar{\underline{d}} = \Sigma \underline{f}_j \underline{d}_j / \underline{n}$, and the summation is over that range of $\underline{x}$ observed across samples. The values in Table 11 were computed from the data in Tables 9 and 10, summing over verbal raw scores 1 to 80 and mathematical raw scores -8 to 55 using the corresponding frequencies for the total group taking Form

V4 when it was first administered in December 1973. The results are directly comparable with those in the SAT scale stability study.

Figure 19 depicts the weighted mean squared difference in terms of its two components. The curved lines in the figure represent four levels of weighted mean squared error: 25, 100, 225, and 400. A particular point on a line is equal to the standard deviation of the difference squared plus the mean difference squared. The equating models are represented by numbers in the case of verbal equatings and by letters in the case of mathematical equatings.

- - - - - - - - - - - - - - - -

Insert Figure 19 about here

- - - - - - - - - - - - - - - -

Comparisons of Equating Results - Variations of the Approximate Three-Parameter Model

The two primary variants of the approximate three-parameter logistic equating model were based on two groupings of examinees: fifths and twentieths. It was expected that the item parameter estimates based on twentieths would be more accurate and that the equating results would also be more accurate. It is clear from the analysis of the item parameter estimates that the estimates based on twentieths were more accurate, but Figures 16 and 17 show that this increased accuracy carried over only slightly to the equating results. In fact, in some parts of the score range the results based on fifths were more accurate; and, in the case of SAT-mathematical, the results based on fifths were superior to those based on twentieths (see Figure 17 and Table 11). The total mean squared error, however, was small for the approximate three-parameter logistic equating models. Thus, the differences in the results are of little practical significance. The

fact that there were only small differences between the models suggests that coarse groupings for obtaining item parameter estimates may be adequate for some equating purposes.

Corrections for coarse grouping were used for the two variations of the approximate three-parameter logistic equating model. The corrected item parameter estimates were applied to the SAT-verbal data, with the intention of using corrected SAT-mathematical estimates if the verbal results indicated that the corrections were useful. The previous comparison of corrected and uncorrected estimates has already indicated that the corrected estimates were less accurate than the uncorrected estimates for SAT-verbal. Nevertheless, the equating results based on corrected estimates were expected to be more accurate than those based on uncorrected estimates. It is clear from Figure 17 that the corrections had little effect on the equating results. If anything, overall equating accuracy decreased (see Table 11). There are several possible explanations as to why the corrections were not effective. First, the Quantile computer program produced relatively accurate estimates of the item parameters, as has already been discussed. Any corrections might simply have added noise to the estimates. Second, the corrections were determined empirically on the basis of only a few data sets and thus might not have been very reliable. Given these results, mathematical results based on corrected item parameters were not obtained.

Neither of the hypotheses regarding the accuracy of equating for approximate three-parameter logistic equating models was confirmed. Making corrections to item parameter estimates and using more than five groupings may be unnecessary when an approximate three-parameter logistic equating model is used with tests that have little form-to-form variation. This does not mean, however, that the

item parameter estimates from Quantile can be used along with those from LOGIST. It is important that the same method be used for estimating item parameters prior to score equating. The possibility of using estimates from different computer programs was beyond the scope of this study and deserves further investigation.

Comparisons of Equating Results - All Three Approximate Models

Figures 16 and 18 and Table 11 give the equating results for the three approximate equating models and their variations along with the results from the concurrent equating model from the SAT scale stability study. Contrary to expectation, the concurrent equating model did not yield the smallest amount of equating error. For both SAT-verbal and SAT-mathematical, concurrent equating had the largest amount of total error except for the modified three-parameter model. Even in comparison with the modified model, concurrent equating yielded more error at the ends of the score range. It is not clear why a more complex model would yield more error, particularly given the large sample sizes.

Both the approximate three-parameter logistic model and the one-parameter logistic model performed better than modified three-parameter model. Primarily because of a general bias (mean difference), the modified three-parameter model produced scores that were up to 20 points too high for SAT-verbal and up to 7 points too high for SAT-mathematical. The approximate three-parameter logistic model had little equating error for either type of equating. The one-parameter logistic model, interestingly enough, yielded the smallest amount of error for equating SAT-mathematical scores but considerably more error than the approximate models for equating SAT-verbal scores. It is clear from Figure 19 that bias accounted for much of the mean squared error associated with the concurrent,

BICAL, and modified three-parameter model verbal equatings and with the modified three-parameter model mathematical equating.

From these results it is clear that none of the hypotheses regarding the comparisons of the three equating models was confirmed for either verbal or mathematical equating. This was true despite the fact that the SAT-mathematical equatings were more accurate than the verbal equatings. It is not obvious why the mathematical equatings had less error, nor why the one-parameter logistic model had the smallest amount of equating error for SAT-mathematical.

Perhaps the most surprising finding of all was the performance of the various approximate equating models, which, because of their simplicity, were expected to yield more equating error than the concurrent model and the modified three-parameter model. These models performed exceedingly well. Is it possible that equating a test to itself, even when it involves a chain of equatings, is biased in some unknown way? Or are the results really valid for test forms that are very similar to one another in content, length, and difficulty?

Follow-up Analysis

Perhaps the most informative research that could have been initiated was to investigate to what extent the results for the approximate three-parameter logistic model were due to the method of transforming the item parameters. For that model existing score equating parameters for converting raw scores to scaled scores were used to compute the transformation for equating item parameter estimates. As a follow-up analysis, this same procedure was used with the one-parameter logistic item parameters from BICAL and with the three-parameter logistic item parameters from LOGIST to see whether this type of item parameter transformation is superior to other methods of placing item parameter estimates

on a common scale. It is possible that the variation in the common item sections used for item calibration, for example, adds error to both the item parameter estimation and transformation processes. Linking item parameter estimates by use of existing score equating results, if satisfactory, would greatly simplify the problem of creating large pools of items with parameters on a common scale.

Item parameter estimates from BICAL and LOGIST were available on all of the data sets shown in Figure 2. Previously, the item parameters from LOGIST did not have to be transformed to a common scale because the items from a given pair of SAT forms and their common equating test were calibrated together. Thus, they were automatically on the same scale and did not have to be transformed. The item parameters from BICAL, however, had been transformed to a common scale through common items by use of the item transformation procedure built into BICAL. The follow-up analysis involved using score equating information to derive transformations for the LOGIST and BICAL item parameter estimates that existed on the separate data sets used in the Petersen, Cook, and Stocking (1983) study. These were V4-fe, X2-fm, Y3-fw, B3-fk, Y2-fu, Z5-et, and V4-et for SAT-verbal and V4-ff, X2-fm, Y3-fx, B3-fl, Y2-fv, Z5-eu, and V4-eu for SAT-mathematical (see Figure 2).

The method described in the section on the approximate three-parameter logistic model was used to derive item parameter transformations. The resulting transformations are shown in Table 12 along with the item parameter transformations previously derived on common items (equating tests) for BICAL and the three-parameter logistic model. The method used to obtain the common-item transformations for the latter model was the characteristic curve method (Stocking & Lord, 1983).

- - - - - - - - - - - - - - -

Insert Table 12 about here

- - - - - - - - - - - - - - -

The results of the equating using the item parameter transformations based on score equating information are shown in Table 13 and Figures 20, 21, and 22. The codes used in Figures 20, 21, and 22 are as follows:

CRIT:	Criterion -- initial scaled score,
CONCUR:	Concurrent (three-parameter logistic),
LOGIST(MOD):	LOGIST (three-parameter logistic) -- modified to use item parameter transformations derived from score equating information,
BICAL(MOD):	BICAL (one parameter logistic) -- modified to use item parameter transformations derived from score equating information, and
W05:	Approximate three-parameter logistic based on fifths without corrections to item parameter estimates.

- -

Insert Table 13 and Figures 20, 21, and 22 about here

- -

The results of the reanalysis show clearly that the use of score equating information to transform item parameters explains why the approximate three-parameter models had a small amount of mean squared error. Recall that the use of score equating information to derive item parameter transformations was built into the approximate three-parameter procedure. A comparison of Tables 11 and 13

indicates that in the case of SAT-verbal the mean squared error decreased from 83.35 to 1.69 for the one-parameter model (BICAL vs. BICAL (modified)) and from 125.15 to 6.99 for the three-parameter model (concurrent vs. LOGIST (modified)). Large decreases in the mean squared error for SAT-mathematical are also evident for the three-parameter model. One can infer that the use of item parameter transformations derived from score equating information was very effective in reducing mean squared error. One can also infer that the effectiveness of the approximate three-parameter models was due to the use of score equating information for deriving transformations.

The reduction in mean squared error is such that the differences among the models utilizing score equating information is slight. However, one of the approximate models, in this case BICAL or BICAL (modified), was still the best model for either SAT-verbal or SAT-mathematical.

The use of item parameter transformations based on score equating information needs further study before they can be applied operationally in testing programs. In this study the chain was short, and half of the common-item linkages used in the study were the very same ones that had been used in score equating. This probably created a situation in which the conversion parameters were more consistent with item calibration results than would be expected if items were calibrated for test forms widely separated in the genealogical chart. One needs to find out how well these transformations work when items from a variety of old forms, particularly those linked together by long equating chains, are calibrated. In conclusion, item parameter transformations based on score equating information look promising, but need further testing.

Recommendations for Further Research

Because of these unusual results, further research is recommended. One possibility is to choose base forms other than Form V4 and redo the chain equating. If the same results were obtained, the findings would be more generalizable, and chance compensating effects at intermediate steps could be ruled out as a possible explanation for the results.

One might also create a situation, as Marco, Petersen, and Stewart (1983) did, in which a test is equated to a different test rather than to itself. One could equate scores on a particular form to scores on the form that is next to it in the original chain equating by proceeding both ways around the circle. For example, scores on Form Z5 could be equated to Form V4 scores by two different paths. The results from these two equatings should agree if the equating method is working properly. Unfortunately, this type of equating is not entirely definitive, because results for two different equatings might agree well but still not be correct.

Further, as was suggested in the previous section, further evaluation is needed of using score equating information to derive item parameter transformations. In the current study the number of links on the chain was limited. This could have created a situation that was favorable to using score equating information for transforming item parameters. The usefulness of the method should be evaluated in situations where score equating is relatively independent of the forms used in the experimental chain.

Finally, and perhaps most important, designs for evaluating equating should be studied under simulated conditions where the correct results are known. Ideally, the simulated conditions should not be based on any of the models for equating. Some useful information could, however, be derived from studies using

the three-parameter logistic model to generate the data. However, other models should also be used to generate data. The use of simulated data would allow one to evaluate any bias that may be created when a test is equated to itself. It would also allow one to evaluate the various equating models in situations in which a test is equated to a different test. It is critical that the usefulness of the current design be evaluated so that decision-makers have a better basis for choosing equating models.

The Air Force Human Resources Laboratory has recently issued a report that reviews various methods of equating mental tests, including IRT models (Gialluca, Crichton, & Vale, 1984). This report and the IRT studies cited, plus a review of other research that has been conducted at ETS and elsewhere, should guide future research activities.

Reference Notes

1. Marco, G. L., Douglass, J. B., & Wingersky, M. S. (August 1982). Approximate item response theory methods for equating test scores. In W. H. Angoff (Chair), Recent developments in equating at ETS. Symposium presented at the meeting of the American Psychological Association, Washington, DC.

2. Wright, B. D., & Mead, R. J. (1976). BICAL: Calibrating items with the Rasch model (Research Memorandum No. 23). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

References

Bock, R. D. (1976). Basic issues in the measurement of change (Appendix). In D. N. M. DeGruijter & L. J. T. Van der Kamp (Eds.), Advances in psychological and educational measurement. London: Wiley.

Douglass, J. B. (1980). A comparison of item characteristic curve models for a classroom examination system. Unpublished doctoral dissertation, Michigan State University.

Gialluca, K. A., Crichton, L. I., & Vale, C. D. (November 1984). Methods for equating mental tests (AFHRL-RT-84-35). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. Journal of Educational Measurement, 19, 139-147.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 18, 1-11.

Levine, R. S. (1955). Equating the score scales of alternate forms administered to samples of different ability (Research Bulletin 55-23). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. Journal of Educational Measurement, 17, 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 14, 139-160.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating methods. In David J. Weiss (Ed.), New horizons in testing. New York: Academic Press.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). Item response theory versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 137-156.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Ree, J. M. (1979). Estimating item characteristic curves. Applied Psychological Measurement, 3, 371-385.

Rentz, R. R., & Bashaw, W. L. (September 1975). Equating reading tests with the Rasch model (2 vols.). Athens, GA: University of Georgia, Educational Research Laboratory, College of Education.

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 15, 23-35.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Wright, B. D., & Stone, M. (1979). Best test design: Rasch measurement. Chicago: MESA.

Footnote

[1]This study was supported by ETS research and development funds provided through the Program Research Planning Council. The data for the study were the scored item tapes used by Petersen, Cook, and Stocking (1983) and derived from records of the College Board Admissions Testing Program, administered by ETS. The authors wish to thank Edwin O. Blew, Martha L. Stocking, and Karen Carroll for processing and analyzing the data used in the study. A preliminary version of this report was presented at the meeting of the American Psychological Association in 1982 (Marco, Douglass, & Wingersky, Note 1).

Table 1

Summary Statistics for SAT-Verbal and SAT-Mathematical Equating Samples

		Verbal				Mathematical			
				Scaled Score[a]				Scaled Score[a]	
Form	Admin. Date	Equating Test	N	Mean	SD	Equating Test	N	Mean	SD
V4	12/73	fe	2665	438	114	ff	2628	457	115
X2	4/75	fe	2686	437	106	ff	2629	476	111
X2	4/75	fm	2562	432	106	fn	2527	471	111
Y3	6/76	fm	2578	426	112	fn	2553	465	113
Y3	1/78	fw	2549	405	109	fx	2455	443	117
B3	5/79	fw	2700	433	108	fx	2633	479	114
B3	5/79	fk	2665	429	104	f1	2596	476	111
Y2	4/76	fk	2879	432	108	f1	2815	469	115
Y2	4/76	fu	2774	428	105	fv	2721	472	115
Z5	12/77	fu	2853	414	108	fv	2774	447	114
Z5	12/77	et	2814	417	110	eu	2739	444	113
V4	12/73	et	2670	436	113	eu	2673	455	115

[a]Scaled score statistics are linear transformations of raw score statistics and deviate slightly from reported score statistics in those cases where curvilinear transformations were used operationally.

Table 2

A Description of Three Approximate IRT Equating Models

Dimension	One-Parameter Logistic (Rasch)	Approximate Three-Parameter Logistic	Modified Three-Parameter Logistic
Type of Data	One data set for each pair of equating tests and SAT forms	One data set for each SAT form	One data set for each pair of equating tests and SAT forms
Method of Item Calibration	BICAL	Quantile (modified LOGIST)	LOGIST
Method of Item Parameter Transformation	Equating of b's	Equating of $\hat{\theta}$'s	Concurrent Calibration

Table 3

Transformations for Equating Item Difficulties Estimated from Different BICAL Item Calibrations

Scale Relationship	Verbal	Mathematical
X2 to V4	$\underline{b}$ - .090	$\underline{b}$ - .131
Y3 to V4	$\underline{b}$ - .174	$\underline{b}$ - .226
B3 to V4	$\underline{b}$ - .214	$\underline{b}$ - .182
Y2 to V4	$\underline{b}$ - .130	$\underline{b}$ - .189
Z5 to V4	$\underline{b}$ - .130	$\underline{b}$ - .116
V4 to V4	$\underline{b}$ - .122	$\underline{b}$ - .014

Table 4

Corrections Applied to Quantile Estimates of Item Parameters

Parameter	Fifths	Twentieths
	SAT-Verbal	
a	$1.184a - .047$	$1.168a - .084$
b	$.965b + .042$	$1.018b + .028$
c estimated	$.935c + .015$	$1.020c + .008$
c common	$c + .006$	$c + .001$
	SAT-Mathematical	
Four-choice items:		
a	$1.149a - .065$	$1.051a - .038$
b	$1.054b + .021$	$1.062b + .042$
c estimated	$1.046c - .008$	$1.080c - .007$
c common	$c - .034$	$c - .001$
Five-choice items:		
a	$1.079a - .011$	$1.038a - .031$
b	$1.021b + .034$	$1.039b + .028$
c estimated	$1.037c - .012$	$1.065c - .009$
c common	$c - .031$	$c - .023$

Table 5

Scaling Parameters for SAT-Verbal and SAT-Mathematical Forms[a]

	Verbal		Mathematical	
SAT Form	A	B	A	B
V4	6.9931	193.2421	8.8839	272.5925
X2	6.9304	193.0270	8.6393	265.4098
Y3	6.8813	189.1449	8.4892	260.3593
B3	6.8315	183.8779	8.5734	267.4198
Y2	7.2588	184.5344	8.5533	269.7831
Z5	6.9440	200.4501	8.4740	273.1202

[a]The 200 to 800 College Board scaled score (S) is determined by the formula $S = AX + B$, where X is the raw score (number right corrected for guessing).

Table 6

Transformations for Equating Item Difficulties Estimated from Different Item Calibrations Using the Computer Program Quantile[a]

Scale Relationship	Uncorrected		Corrected	
	Fifths	Twentieths	Fifths	Twentieths
		SAT-Verbal		
X2 to V4	.938b - .049	.944b - .055	.942b - .044	.950b - .050
Y3 to V4	1.013b - .326	1.024b - .319	1.017b - .311	1.033b - .323
B3 to V4	.916b - .068	.915b - .073	.915b - .054	.921b - .071
Y2 to V4	.925b - .081	.926b - .085	.924b - .073	.928b - .075
Z5 to V4	1.011b - .214	1.020b - .218	1.016b - .202	1.034b - .216
V4 to V4	.994b - .018	.991b - .014	.991b - .019	.991b - .012
		SAT-Mathematical		
X2 to V4	.986b + .064	.997b + .060		
Y3 to V4	1.127b - .243	1.121b - .231		
B3 to V4	.985b + .109	.999b + .100	Not Determined	
Y2 to V4	1.039b + .058	1.073b + .042		
Z5 to V4	1.097b - .233	1.099b - .231		
V4 to V4	1.010b - .016	1.015b - .017		

[a]The formula for transforming item discrimination parameters to the scale for Form V4 is a/A, where A is the slope parameter given in the table.

Table 7

Comparison of Quantile Results to True Values and LOGIST Results for Artificial Data

	True Values	LOGIST Results	Relative to True Values: Quantiles 5ths	Relative to True Values: Quantiles 20ths	Relative to LOGIST Results: Quantiles 5ths	Relative to LOGIST Results: Quantiles 20ths
No. of items = 45						
Mean Absolute Difference between Item Response Functions[a]		.0177	.0173	.0171	.0169	.0131
a Parameter						
Mean	.917	.975	.821	.879		
Standard Deviation	.328	.362	.306	.308		
Mean Absolute Difference		.124	.156	.124	.183	.122
Root Mean Squared Error		.149	.203	.161	.218	.140
Mean Difference		.058	-.096	-.037	-.154	-.096
SD of Difference		.138	.181	.158	.157	.104
Correlation		.924	.839	.877	.904	.965
b Parameter						
Mean	.202	.201	.190	.156		
Standard Deviation	.987	.993	.984	.907		
Mean Absolute Difference		.106	.153	.144	.128	.118
Root Mean Squared Error		.143	.238	.196	.209	.160
Mean Difference		-.001	-.013	-.046	-.012	-.046
SD of Difference		.144	.240	.192	.211	.155
Correlation		.989	.970	.983	.977	.991
c Parameter						
Mean	.195	.195	.175	.178		
Standard Deviation	.053	.061	.092	.086		
Mean Absolute Difference		.035	.062	.055	.048	.040
Root Mean Squared Error		.043	.082	.076	.073	.060
Mean Difference		-.000	-.020	-.017	-.020	-.017
SD of Difference		.044	.081	.074	.071	.059
Correlation[b]		.716	.482	.512	.639	.732

[a]This is the mean absolute difference between the item response functions averaged over all of the abilities in the criterion group and then averaged over all of the items.

[b]This includes the c 's fixed at the common c value.

Table 8

Comparison of Quantile Results, without Corrections and with Corrections, to LOGIST Results for Form V4

	Relative to LOGIST Results									
		V4fe				V4et				
		w/o Corrections		w/ Corrections			w/o Corrections		w/ Corrections	
	LOGIST	5ths	20ths	5ths	20ths	LOGIST	5ths	20ths	5ths	20ths
No. of items = 90										
Mean Absolute Difference between Item Response Functions[a]		.0090	.0056	.0101	.0074		.0090	.0085	.0122	.0088
a Parameter										
Mean	.782	.748	.776	.839	.823	.762	.757	.785	.850	.833
Standard Deviation	.274	.289	.266	.342	.311	.295	.278	.275	.329	.321
Mean Absolute Difference		.091	.048	.095	.065		.092	.057	.110	.078
Root Mean Squared Error		.149	.084	.180	.107		.143	.082	.181	.192
Mean Difference		-.034	-.006	.057	.041		-.004	.023	.088	.072
SD of Difference		.146	.085	.172	.099		.144	.079	.159	.086
Correlation		.866	.951	.866	.951		.876	.964	.876	.964
b Parameter										
Mean	.304	.277	.256	.309	.288	.326	.303	.276	.335	.309
Standard Deviation	1.313	1.363	1.330	1.315	1.354	1.387	1.374	1.344	1.325	1.368
Mean Absolute Difference		.101	.080	.080	.076		.090	.093	.089	.078
Root Mean Squared Error		.171	.151	.159	.150		.134	.128	.143	.114
Mean Difference		-.027	-.049	.005	-.016		-.023	-.050	.009	-.017
SD of Difference		.170	.144	.159	.150		.133	.119	.144	.113
Correlation		.993	.994	.993	.994		.995	.997	.995	.997
c Parameter										
Mean	.153	.148	.140	.153	.151	.156	.156	.145	.161	.156
Standard Deviation	.060	.053	.055	.050	.056	.050	.056	.063	.053	.064
Mean Absolute Difference		.023	.024	.020	.022		.020	.028	.018	.026
Root Mean Squared Error		.051	.050	.050	.049		.031	.040	.030	.039
Mean Difference		-.005	-.013	.004	-.002		-.008	-.011	.000	-.001
SD of Difference		.051	.049	.050	.049		.031	.039	.029	.039
Correlation[b]		.596	.645	.596	.645		.837	.792	.837	.792

[a]This is the mean absolute difference between item response functions averaged over all of the abilities in the criterion group and then averaged over all of the items.

[b]This includes the $\underline{c}$'s fixed at the common $\underline{c}$ value.

Table 9

Initial and Final Transformations of SAT-Verbal Form V4 Raw Scores to Scaled Scores for Chain Equating

Raw Score	Freq[a]	Initial Scaled Score	Final Scaled Score (Chain Equating)						
					Approximate Three-Parameter				
					Fifths		Twentieths		Modified Three-Parameter
			Concurrent	BICAL	w/o Corr	w/ Corr	w/o Corr	w/ Corr	
90	3	822.62	822.62	822.62	822.62	822.62	822.62	822.62	822.62
89	8	815.63	821.64	816.23	816.90	817.06	815.90	815.78	817.88
88	11	808.63	818.56	809.81	810.65	810.92	809.25	809.10	812.80
87	1	801.64	814.25	803.37	804.17	804.53	802.59	802.43	807.46
86	26	794.65	809.20	796.90	797.54	797.98	795.92	795.75	801.93
85	44	787.66	803.69	790.39	790.81	791.31	789.23	789.07	796.26
84	71	780.66	797.89	783.87	783.99	784.56	782.51	782.37	790.45
83	98	773.67	791.90	777.32	777.11	777.73	775.78	775.64	784.53
82	38	766.68	785.78	770.75	770.18	770.85	769.01	768.88	778.31
81	135	759.68	779.55	764.16	763.22	763.93	762.22	762.09	772.42
80	172	752.69	773.20	757.55	756.24	756.99	755.40	755.27	766.26
79	215	745.70	766.75	750.92	749.25	750.03	748.55	748.42	760.02
78	251	738.70	760.18	744.27	742.24	743.06	741.67	741.53	753.72
77	164	731.71	753.51	737.60	735.23	736.07	734.77	734.62	747.35
76	294	724.72	746.75	730.91	728.21	729.07	727.84	727.68	740.93
75	360	717.72	739.88	724.21	721.18	722.06	720.88	720.70	734.45
74	441	710.73	732.92	717.50	714.14	715.04	713.90	713.71	727.93
73	533	703.74	725.87	710.76	707.10	708.00	706.90	706.70	721.35
72	374	696.75	718.75	704.02	700.05	700.95	699.88	699.66	714.74
71	590	689.75	711.54	697.26	692.99	693.88	692.83	692.60	708.08
70	667	682.76	704.30	690.49	685.91	686.80	685.77	685.51	701.30
69	791	675.77	696.99	683.70	678.83	679.70	678.68	678.41	694.64
68	913	668.77	689.63	676.90	671.73	672.58	671.58	671.28	687.87
67	650	661.78	682.24	670.09	664.62	665.45	664.46	664.14	681.06
66	1002	654.79	674.81	663.27	657.50	658.31	657.31	656.97	674.21
65	1166	647.79	667.34	656.44	650.37	651.15	650.15	649.79	667.34
64	1313	640.80	659.87	649.60	643.23	643.97	642.98	642.59	660.43
63	1486	633.81	652.37	642.74	636.08	636.78	635.79	635.36	653.48
62	1022	626.81	644.86	635.88	628.91	629.58	628.58	628.12	646.52
61	1562	619.82	637.34	629.00	621.74	622.37	621.36	620.87	639.52
60	1801	612.83	629.81	622.11	614.56	615.14	614.13	613.60	632.50
59	1955	605.84	622.26	615.22	607.37	607.90	606.88	606.33	625.46
58	2182	598.84	614.73	608.32	600.16	600.65	599.63	599.04	618.40
57	1633	591.85	607.19	601.41	592.95	593.39	592.37	591.74	611.32
56	2228	584.86	599.68	594.50	585.73	586.11	585.11	584.44	604.22
55	2583	577.86	592.20	587.57	578.50	578.82	577.85	577.14	597.10
54	2694	570.87	584.70	580.63	571.27	571.52	570.59	569.85	589.96
53	2987	563.88	577.24	573.69	564.03	564.21	563.33	562.56	582.81
52	2202	556.88	569.80	566.74	556.78	556.90	556.08	555.28	575.64
51	3133	549.89	562.39	559.79	549.54	549.58	548.83	548.00	568.45
50	3474	542.90	555.01	552.83	542.29	542.26	541.59	540.75	561.26

Table 9 (Continued)

Raw Score	Freq[a]	Initial Scaled Score	Final Scaled Score (Chain Equating)						
					Approximate Three-Parameter				
					Fifths		Twentieths		Modified Three-Parameter
			Concurrent	BICAL	w/o Corr	w/ Corr	w/o Corr	w/ Corr	
49	3653	535.90	547.67	545.85	535.04	534.95	534.37	533.51	554.05
48	3998	528.91	540.36	538.88	527.80	527.63	527.16	526.29	546.83
47	3011	521.92	533.09	531.90	520.56	520.33	519.97	519.09	539.61
46	4194	514.92	525.89	524.91	513.34	513.04	512.80	511.92	532.37
45	4418	507.93	518.73	517.91	506.12	505.77	505.63	504.77	525.13
44	4729	500.94	511.59	510.91	498.92	498.51	498.52	497.66	517.88
43	4848	493.95	504.49	503.91	491.74	491.28	491.41	490.56	510.62
42	3793	486.95	497.44	496.90	484.58	484.08	484.32	483.50	503.36
41	5032	479.96	490.42	489.88	477.44	476.90	477.26	476.47	496.08
40	5449	472.97	483.44	482.86	470.33	469.76	470.23	469.47	488.80
39	5492	465.97	476.48	475.83	463.24	462.64	463.21	462.49	481.52
38	5835	458.98	469.56	468.80	456.18	455.57	456.22	455.54	474.23
37	4420	451.99	462.64	461.77	449.13	448.52	449.25	448.62	466.93
36	5794	444.99	455.74	454.73	442.14	441.51	442.31	441.72	459.63
35	5942	438.00	448.85	447.68	435.16	434.54	435.38	434.84	452.32
34	5933	431.01	441.95	440.63	428.22	427.60	428.48	427.99	445.01
33	6072	424.01	435.04	433.58	421.30	420.69	421.60	421.16	437.69
32	4537	417.02	428.11	426.52	414.41	413.82	414.73	414.35	430.36
31	5860	410.03	421.16	419.46	407.56	406.98	407.89	407.55	423.03
30	5996	403.04	414.16	412.39	400.72	400.17	401.06	400.78	415.70
29	6005	396.04	407.14	405.32	393.92	393.40	394.24	394.02	408.36
28	5990	389.05	400.05	398.25	387.14	386.65	387.45	387.27	401.01
27	4521	382.06	392.93	391.17	380.38	379.92	380.66	380.54	393.66
26	5359	375.06	385.76	384.09	373.65	373.22	373.89	373.82	386.29
25	5381	368.07	378.54	377.01	366.94	366.54	367.14	367.12	378.93
24	5325	361.08	371.25	369.92	360.25	359.87	360.39	360.42	371.56
23	5002	354.08	363.92	362.83	353.57	353.23	353.66	353.74	364.18
22	3719	347.09	356.55	355.74	346.91	346.60	346.93	347.06	356.80
21	4666	340.10	349.10	348.64	340.27	339.98	340.22	340.39	349.42
20	4491	333.10	341.60	341.54	333.63	333.37	333.51	333.73	342.02
19	4356	326.11	334.05	334.43	327.01	326.78	326.82	327.08	334.63
18	4086	319.12	326.44	327.33	320.40	320.18	320.13	320.44	327.23
17	2933	312.12	318.77	320.22	313.79	313.59	313.45	313.80	319.83
16	3472	305.13	311.05	313.10	307.19	307.00	306.77	307.17	312.43
15	3320	298.14	303.29	305.98	300.58	300.41	300.10	300.54	305.03
14	3112	291.15	295.48	298.87	293.98	293.82	293.43	293.91	297.63
13	2836	284.15	287.65	291.74	287.37	287.22	286.76	287.28	290.24
12	1967	277.16	279.80	284.62	280.76	280.61	280.09	280.65	282.85
11	2315	270.17	271.98	277.49	274.14	274.00	273.42	274.01	275.47
10	2209	263.17	264.18	270.35	267.52	267.37	266.74	267.36	268.09
9	1967	256.18	256.45	263.22	260.88	260.72	260.06	260.69	260.73
8	1754	249.19	248.78	256.08	254.23	254.06	253.36	254.01	253.37
7	1120	242.19	241.23	248.94	247.56	247.38	246.65	247.31	246.03
6	1290	235.20	233.81	241.79	240.87	240.68	239.92	240.58	238.71
5	1199	228.21	226.54	234.64	234.16	233.95	233.17	233.83	231.40

Table 9 (Continued)

Raw Score	Freq[a]	Initial Scaled Score	Final Scaled Score (Chain Equating)						
					Approximate Three-Parameter				
					Fifths		Twentieths		Modified Three-Parameter
			Concurrent	BICAL	w/o Corr	w/ Corr	w/o Corr	w/ Corr	
4	1057	221.21	219.44	227.49	227.42	227.20	226.39	227.04	224.11
3	853	214.22	212.52	220.33	220.65	220.41	219.59	220.20	216.83
2	484	207.23	205.78	213.17	213.85	213.59	212.75	213.32	209.58
1	522	200.24	199.22	206.01	207.01	206.71	205.87	206.39	202.34
0	484	193.24	192.84	198.84	200.12	199.80	198.94	199.39	195.12
-1	358	186.25	186.63	191.67	193.19	192.82	191.96	192.32	187.92
-2	265	179.26	180.59	184.49	186.20	185.79	184.91	185.16	180.74
-3	112	172.26	174.70	177.31	179.15	178.69	177.80	177.90	173.58
-4	129	165.27	168.97	170.13	172.03	171.53	170.61	170.52	166.44
-5	91	158.28	163.67	162.94	164.82	164.21	163.31	162.93	159.36
-6	52	151.28	156.46	155.74	157.36	156.89	155.87	155.29	152.51
-7	30	144.29	149.24	148.54	150.23	149.77	148.23	148.23	145.50
-8	6	137.30	142.01	141.34	143.09	142.65	141.16	141.16	138.47
-9	14	130.30	134.79	134.12	135.96	135.53	134.09	134.10	131.45
-10	6	123.31	127.56	126.91	128.82	128.41	127.02	127.03	124.42
-11	3	116.32	120.34	119.68	121.69	121.29	119.95	119.97	117.39
-12	2	109.32	113.11	112.45	114.55	114.17	112.88	112.91	110.37
-13	0	102.33	105.89	105.21	107.42	107.05	105.81	105.84	103.34
-14	1	95.34	98.66	97.97	100.28	99.93	98.74	98.78	96.31
-15	0	88.35	91.44	90.71	93.15	92.81	91.67	91.71	89.29
-16	0	81.35	84.21	83.45	86.01	85.69	84.59	84.65	82.26
-17	0	74.36	76.99	76.17	78.88	78.57	77.52	77.59	75.24
-18	0	67.37	69.76	68.88	71.74	71.45	70.45	70.52	68.21
-19	0	60.37	62.54	61.58	64.61	64.33	63.38	63.46	61.18
-20	0	53.38	55.31	54.26	57.47	57.21	56.31	56.39	54.16
-21	0	46.39	48.09	46.93	50.34	50.09	49.24	49.33	47.13
-22	0	39.39	40.86	39.39	43.20	42.97	42.17	42.27	40.10

[a] SAT-verbal form V4 raw score frequency distribution for initial December 1973 administration.

Table 10

Initial and Final Transformations of SAT-Mathematical Form V4 Raw Scores to Scaled Scores for Chain Equating

Raw Score	Freq[a]	Initial Scaled Score	Final Scaled Score (Chain Equating)				
			Concurrent	BICAL	Approximate Three-Parameter		Modified Three-Parameter
					Fifths[b]	Twentieths[b]	
60	34	805.63	805.63	805.63	805.63	805.63	805.63
59	114	796.74	799.30	796.88	796.06	797.41	798.28
58	205	787.86	791.93	788.07	787.01	788.92	789.85
57	47	778.97	784.12	779.22	778.16	780.41	781.12
56	317	770.09	776.15	770.34	769.44	771.90	772.32
55	475	761.21	768.13	761.44	760.79	763.36	763.56
54	638	752.32	760.07	752.52	752.19	754.82	754.87
53	795	743.44	751.93	743.59	743.61	746.23	746.25
52	430	734.56	743.70	734.66	735.03	737.61	737.68
51	878	725.67	735.35	725.73	726.44	728.94	729.15
50	1094	716.79	726.86	716.79	717.83	720.23	720.64
49	1312	707.90	718.26	707.85	709.20	711.46	712.14
48	1555	699.02	709.52	698.91	700.53	702.66	703.64
47	1144	690.14	700.66	689.97	691.83	693.80	695.13
46	1647	681.25	691.67	681.04	683.08	684.89	686.59
45	1939	672.37	682.57	672.11	674.30	675.94	678.02
44	2235	663.48	673.37	663.18	665.47	666.93	669.40
43	2452	654.60	664.06	654.25	656.59	657.88	660.73
42	2117	645.72	654.68	645.32	647.67	648.79	652.02
41	2588	636.83	645.24	636.39	638.72	639.65	643.26
40	2859	627.95	635.74	627.47	629.71	630.47	634.45
39	3213	619.06	626.23	618.54	620.67	621.25	625.60
38	3560	610.18	616.68	609.63	611.59	612.01	616.71
37	3059	601.30	607.14	600.71	602.48	602.74	607.79
36	3838	592.41	597.61	591.79	593.35	593.45	598.83
35	4022	583.53	588.10	582.88	584.20	584.16	589.85
34	4385	574.65	578.64	573.97	575.04	574.87	580.86
33	4761	565.76	569.23	565.06	565.88	565.59	571.85
32	3817	556.88	559.89	556.15	556.72	556.33	562.84
31	5053	547.99	550.62	547.25	547.57	547.10	553.83
30	5248	539.11	541.45	538.35	538.44	537.90	544.81
29	5511	530.23	532.36	529.45	529.34	528.73	535.81
28	5870	521.34	523.36	520.56	520.26	519.61	526.83
27	4779	512.46	514.46	511.66	511.20	510.54	517.84
26	6034	503.57	505.63	502.77	502.19	501.51	508.88
25	6466	494.69	496.88	493.89	493.20	492.52	499.93
24	6576	485.81	488.19	485.00	484.24	483.58	490.99
23	6894	476.92	479.55	476.12	475.32	474.67	482.07
22	5470	468.04	470.95	467.24	466.43	465.80	473.16
21	6934	459.15	462.36	458.36	457.56	456.96	464.27
20	6844	450.27	453.78	449.48	448.72	448.16	455.39

Table 10 (Continued)

Raw Score	Freq[a]	Initial Scaled Score	Final Scaled Score (Chain Equating)				
			Concurrent	BICAL	Approximate Three-Parameter		Modified Three-Parameter
					Fifths[b]	Twentieths[b]	
19	7186	441.39	445.20	440.60	439.89	439.38	446.52
18	7208	432.50	436.59	431.73	431.09	430.63	437.66
17	5444	423.62	427.95	422.86	422.30	421.89	428.80
16	6909	414.73	419.27	413.98	413.53	413.17	419.96
15	6986	405.85	410.54	405.11	404.77	404.47	411.10
14	7014	396.97	401.74	396.24	396.02	395.79	402.25
13	6696	388.08	392.88	387.37	387.28	387.11	393.40
12	5006	379.20	383.95	378.50	378.55	378.45	384.54
11	6025	370.32	374.92	369.63	369.83	369.80	375.67
10	5948	361.43	365.79	360.76	361.13	361.16	366.78
9	5588	352.55	356.53	351.90	352.43	352.54	357.87
8	5411	343.66	347.13	343.03	343.75	343.93	348.94
7	3576	334.78	337.57	334.16	335.08	335.34	339.99
6	4577	325.90	327.82	325.30	326.42	326.77	331.00
5	4155	317.01	317.87	316.43	317.78	318.23	321.97
4	3849	308.13	307.71	307.57	309.15	309.70	312.90
3	3405	299.24	297.38	298.70	300.53	301.21	303.78
2	2065	290.36	286.91	289.84	291.93	292.74	294.60
1	2448	281.48	276.39	280.98	283.34	284.31	285.35
0	2049	272.59	265.95	272.13	274.78	275.91	276.02
-1	1621	263.71	255.74	263.27	266.24	267.57	266.60
-2	1152	254.82	245.88	254.41	257.74	259.30	257.08
-3	456	245.94	236.44	245.56	249.26	251.09	247.44
-4	500	237.06	227.18	236.71	240.78	242.96	237.65
-5	324	228.17	216.09	227.86	232.17	234.86	227.70
-6	186	219.29	211.51	219.01	222.96	226.30	217.56
-7	97	210.41	203.07	210.16	213.91	217.06	208.69
-8	17	201.52	194.64	201.32	204.87	207.83	201.70
-9	17	192.64	186.20	192.47	195.82	198.59	192.80
-10	4	183.75	177.76	183.62	186.78	189.35	183.91
-11	0	174.87	169.33	174.78	177.73	180.11	175.02
-12	0	165.99	160.89	165.93	168.69	170.87	166.12
-13	0	157.10	152.46	157.07	159.64	161.64	157.23
-14	0	148.22	144.02	148.21	150.60	152.40	148.33
-15	0	139.33	135.58	139.33	141.55	143.16	139.44

[a] SAT-mathematical Form V4 raw score frequency distribution for initial December 1973 administration.

[b] No correction was used.

Table 11

Information and Summary Discrepancy Indices[a] for Item Response Theory Equating Models

Index	Initial Scale (Criterion)	Concurrent	BICAL	Approximate Three-Parameter				Modified Three-Parameter
				Fifths w/o Corrections	Fifths w/ Corrections	Twentieths w/o Corrections	Twentieths w/ Corrections	
SAT-verbal								
Scaled Score:								
Mean	435.37	445.73	444.45	434.93	434.67	434.75	434.45	448.72
Standard Deviation	109.09	112.89	109.60	108.58	108.79	108.54	108.17	113.38
Mean Squared Error[b]		125.15	83.35	5.75	7.38	4.83	7.01	198.51
Mean Difference		10.36	9.08	-.44	-.70	-.62	-.92	13.35
SD of Difference		4.23	.96	2.36	2.62	2.11	2.48	4.49
SAT-mathematical								
Scaled Score:								
Mean	468.05	471.61	467.40	467.81		467.82		473.35
Standard Deviation	113.25	115.50	113.32	113.43		113.59		113.63
Mean Squared Error[b]		23.27	.46	1.61	Not	3.78	Not	28.67
Mean Difference		3.55	-.66	-.24	Determined	-.23	Determined	5.29
SD of Difference		3.26	.18	1.25		1.93		.81

[a]Computed for SAT-verbal raw scores 1 through 80 and for SAT-mathematical raw scores -8 through 55.

[b](SD of Difference)^2 + (Mean Difference)^2.

Table 12

Transformations for Equating Item Difficulties Estimated from Different Item Calibrations Using BICAL (One-Parameter Model) and LOGIST (Three-Parameter Model)

Scale Relationship	One-Parameter		Three-Parameter	
	Common-Item	Score-Conversion	Common-Item	Score-Conversion
		SAT-Verbal		
X2 - V4	b - .090	.877b - .175	.898b - .002	.931b - .038
Y3 - V4	b - .174	.861b - .189	.957b - .297	.995b - .312
B3 - V4	b - .214	.851b - .177	.839b - .074	.897b - .065
Y2 - V4	b - .130	.906b - .166	.849b - .087	.910b - .082
Z5 - V4	b - .130	.876b - .065	.903b - .231	.997b - .184
V4 - V4	b - .122	1.014b - .046	.898b - .092	.979b + .002
		SAT-Mathematical		
X2 - V4	b - .131	.996b - .159	.948b + .096	.993b + .066
Y3 - V4	b - .226	.950b - .205	1.068b - .211	1.109b - .215
B3 - V4	b - .182	1.021b - .074	.942b + .064	1.005b + .106
Y2 - V4	b - .189	.956b - .113	.959b + .031	1.048b + .058
Z5 - V4	b - .116	.971b - .101	.999b - .259	1.051b - .224
V4 - V4	b - .014	1.010b - .013	.998b - .083	1.010b - .027

Table 13

Information and Summary Discrepancy Indices[a] for Selected Item Response Theory Equating Models

Index	Initial Scale (Criterion)	Concurrent	BICAL	Characteristic Curve Transformation	LOGIST (Modified)	BICAL (Modified)
			SAT-Verbal			
Scaled Score:						
Mean	435.37	445.73	444.45	446.34	434.50	434.33
Standard Deviation	109.09	112.89	109.60	116.47	106.61	108.33
Mean Squared Error[b]		125.15	83.35	178.98	6.99	1.69
Mean Difference		10.36	9.08	10.97	-.87	-1.04
S.D. of Difference		4.23	.96	7.66	2.50	.79
			SAT-Mathematical			
Scaled Score:						
Mean	468.05	471.61	467.40	474.95	468.49	467.39
Standard Deviation	113.25	115.50	113.32	114.82	112.87	112.36
Mean Squared Error[b]		23.27	.46	61.08	7.28	1.31
Mean Difference		3.55	-.66	6.90	.44	-.66
S.D. of Difference		3.26	.18	3.67	2.66	.93

[a]Computed for SAT-verbal raw scores 1 through 80 and for SAT-mathematical raw scores -8 through 55.

[b]$(\text{SD of Difference})^2 + (\text{Mean Difference})^2$.

Six Verbal Equatings

V4 → fe → X2 → fm → Y3 → fw
↑ ↓
et ← Z5 ← fu ← Y2 ← fk ← B3

Six Mathematical Equatings

V4 → ff → X2 → fn → Y3 → fx
↑ ↓
eu ← Z5 ← fv ← Y2 ← f1 ← B3

Figure 1. Verbal and mathematical equating chains.

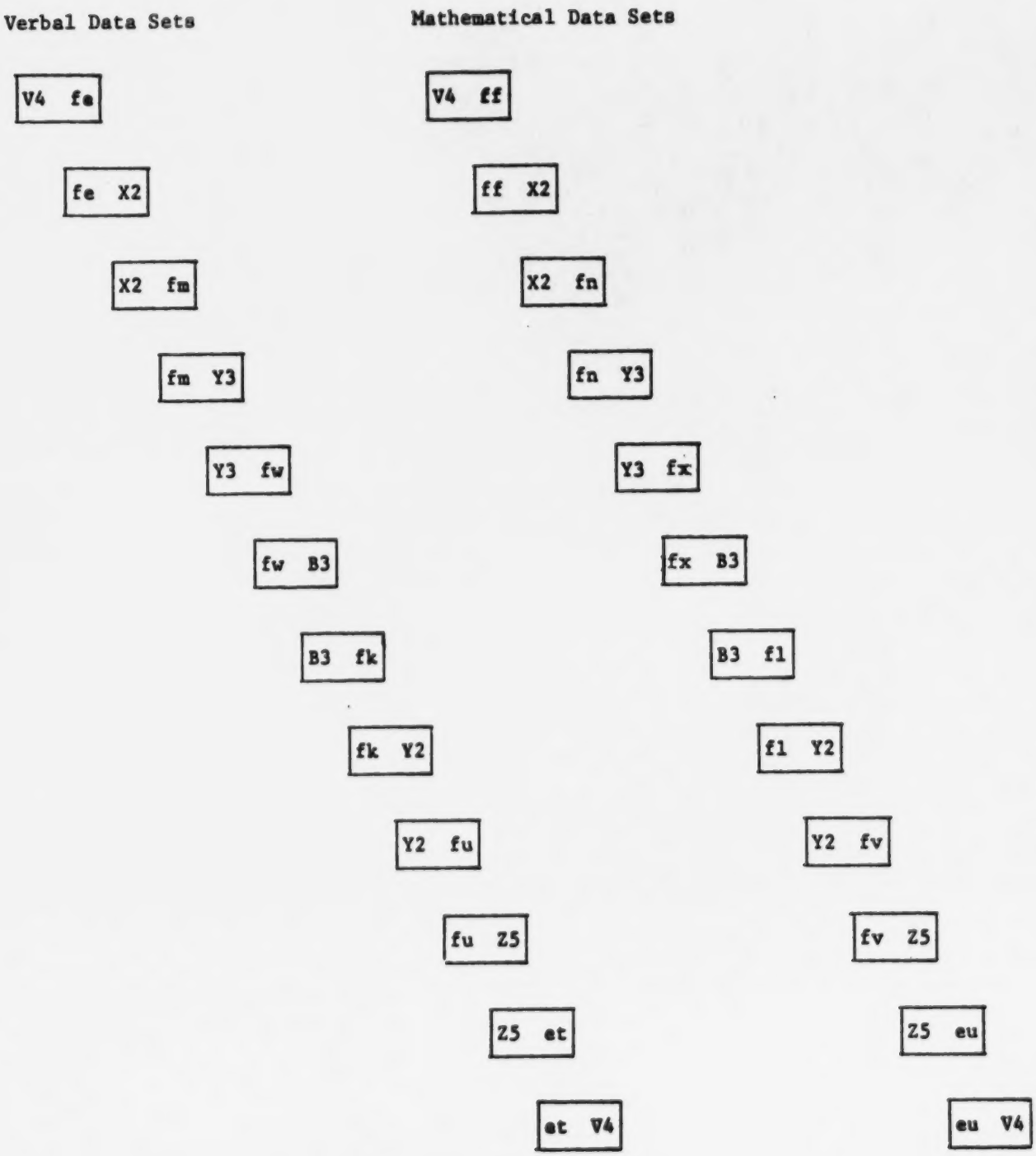
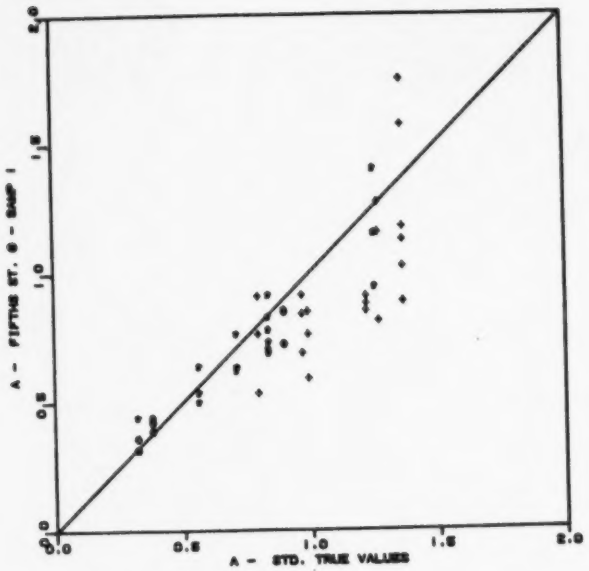
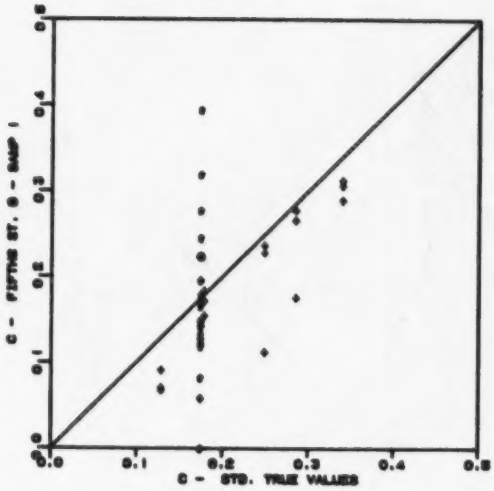
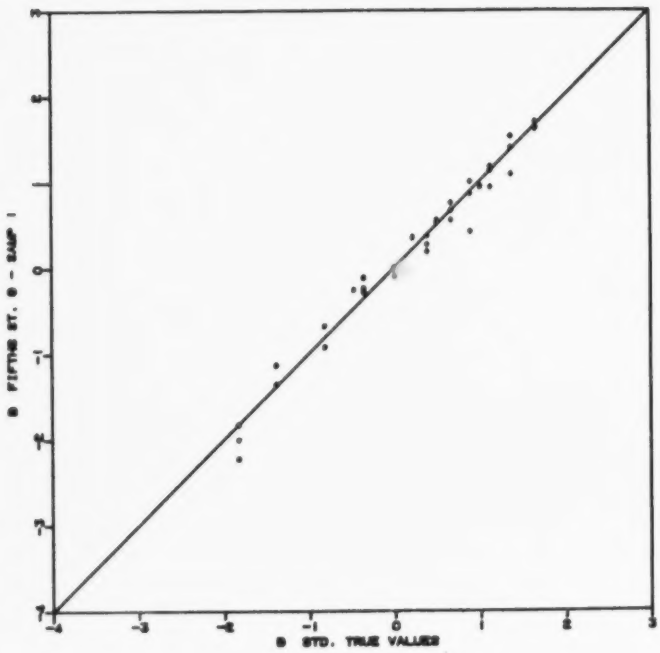


Figure 2. Verbal and mathematical data sets. Each box represents a sample of approximately 2,670 cases.





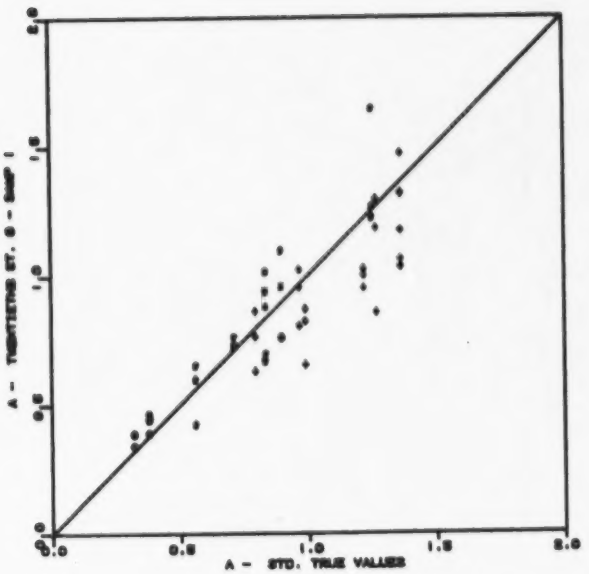


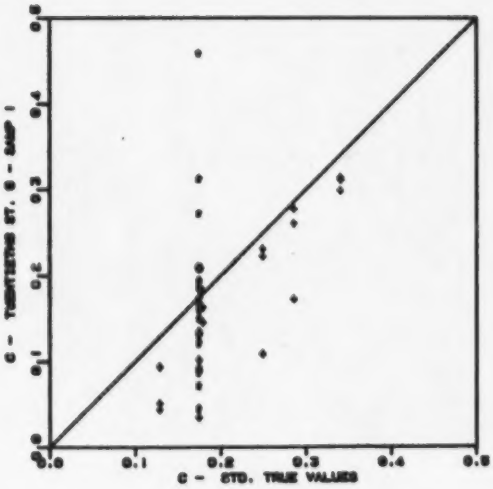
Legend

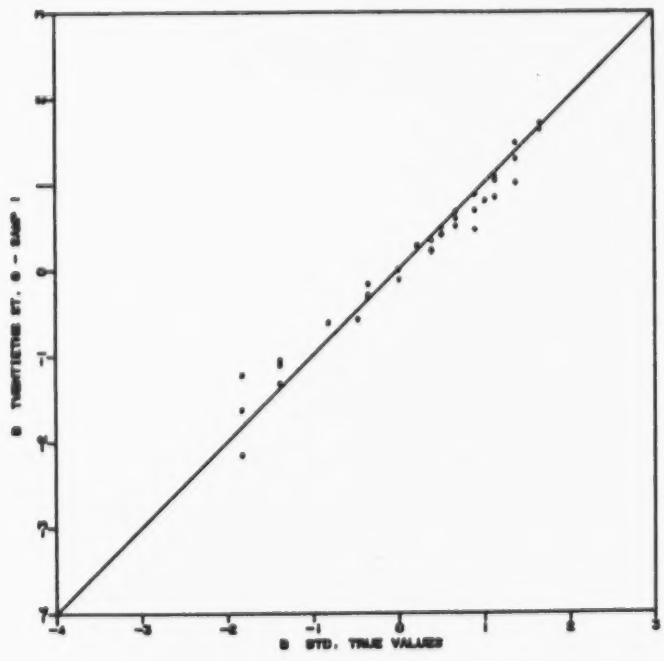
* - one or both c's were fixed at the common c value

+ or # - neither c was fixed at the common c value

Figure 3. Comparison of Quantile parameter estimates for fifths to true values. Artificial data.





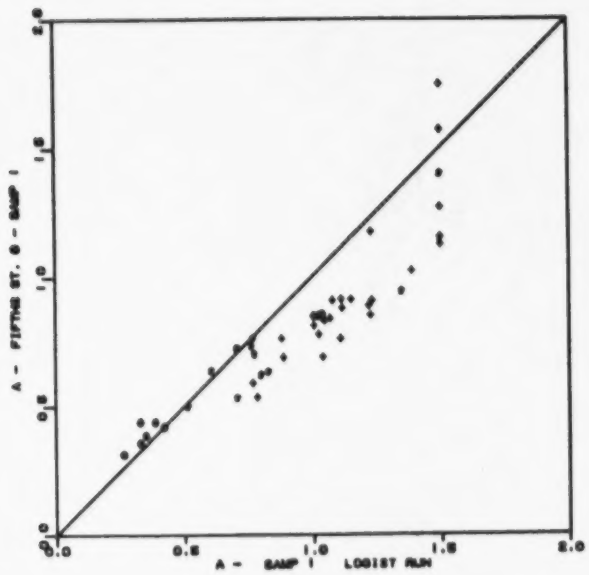


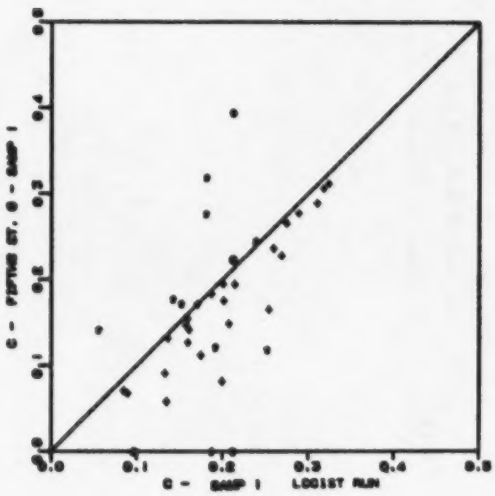
Legend

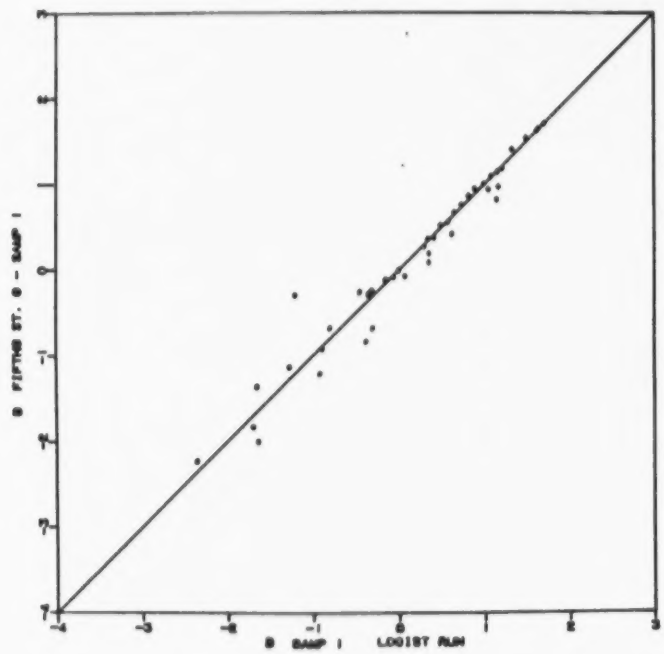
° - one or both c's were fixed at the common c value

+ or * - neither c was fixed at the common c value

Figure 4. Comparison of Quantiles parameter estimates for twentieths to true values. Artificial data.





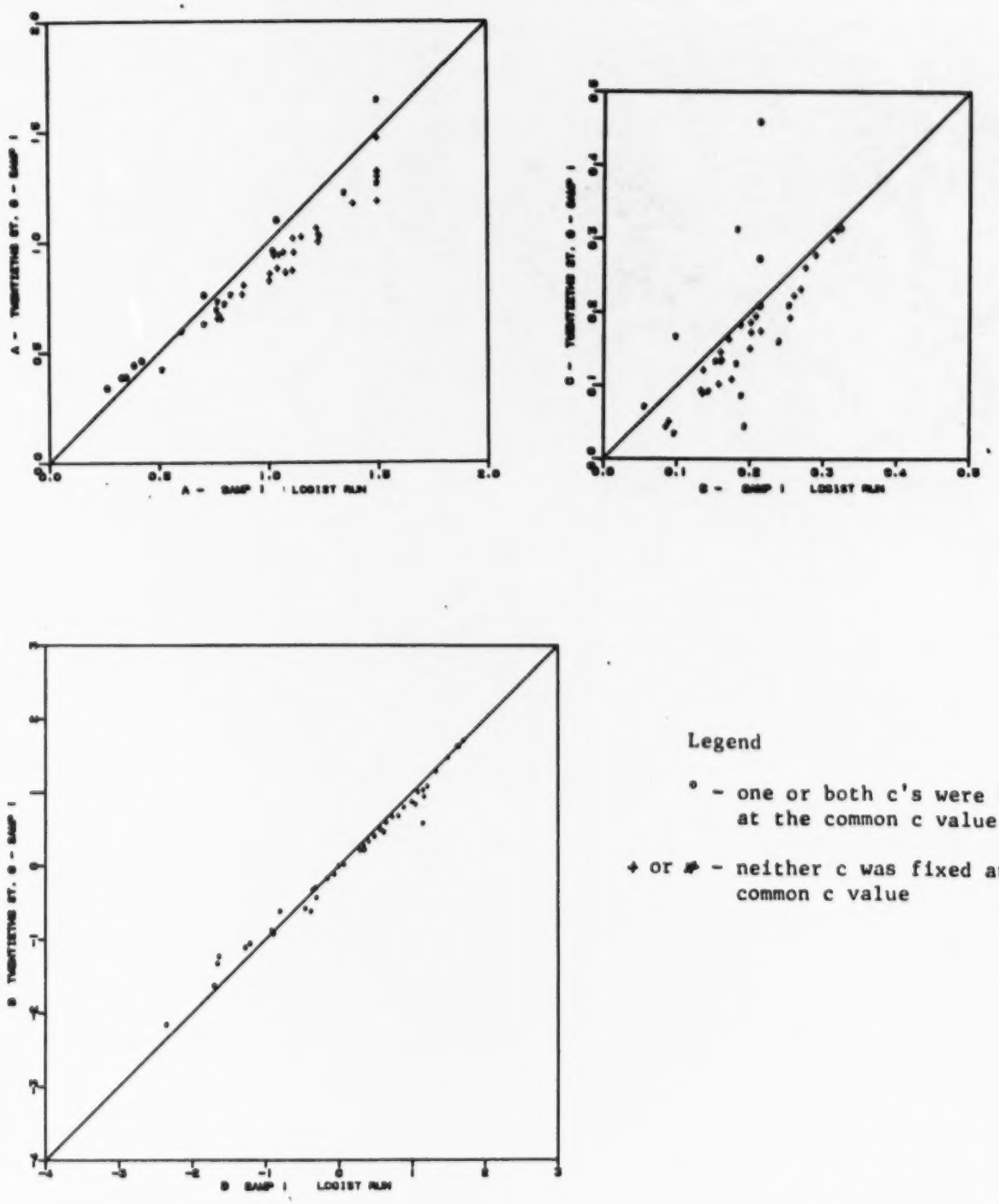


Legend

° - one or both c's were fixed at the common c value

+ or * - neither c was fixed at the common c value

Figure 5. Comparison of Quantile parameter estimates for fifths to LOGIST parameter estimates. Artificial data.

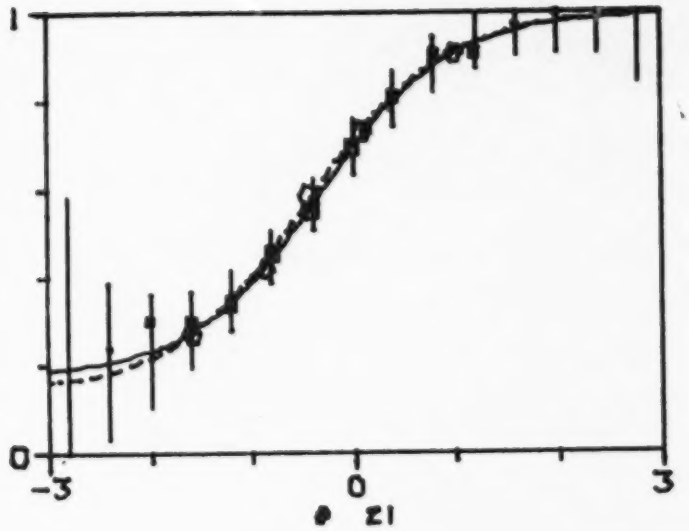


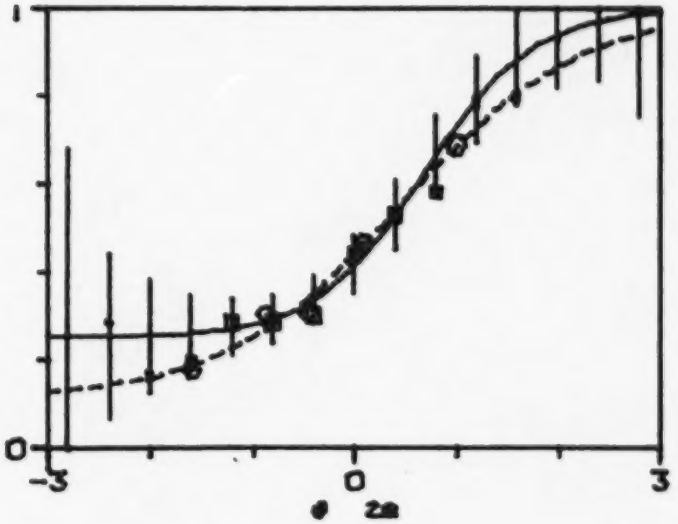
Legend

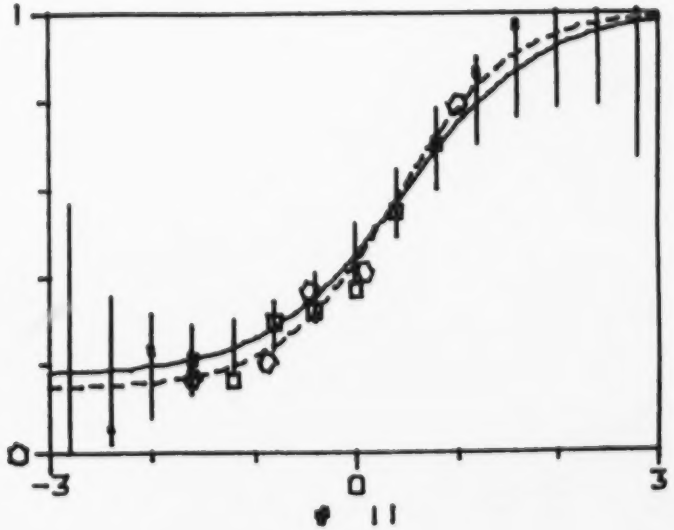
° - one or both c's were fixed at the common c value

+ or * - neither c was fixed at the common c value

Figure 6 . Comparison of Quantile parameter estimates for twentieths to LOGIST parameter estimates. Artificial data.







Legend

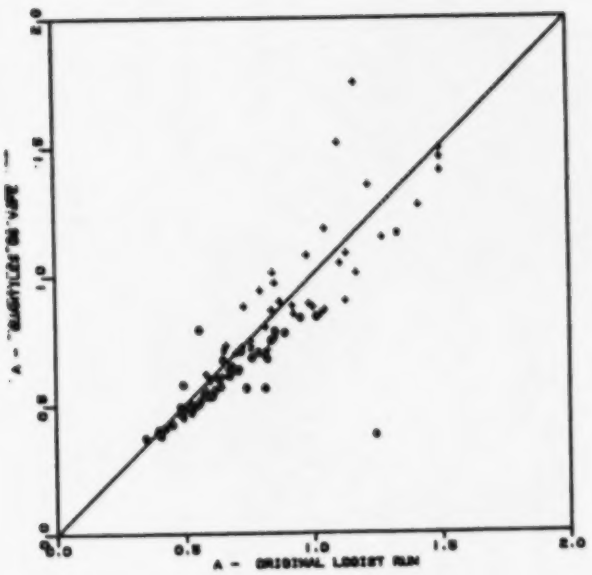
——- Item response function using the true values of the parameters

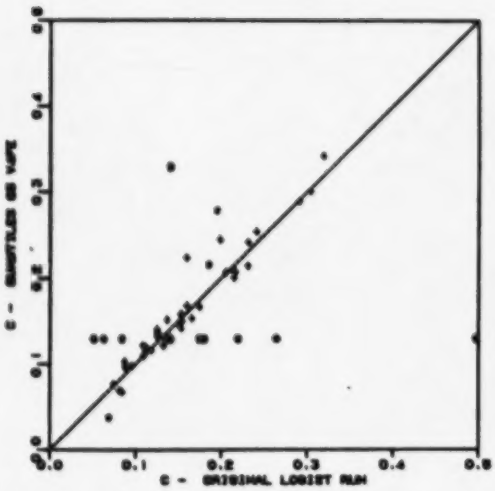
---- Item response function using the quantile estimated parameters for the 5ths grouping

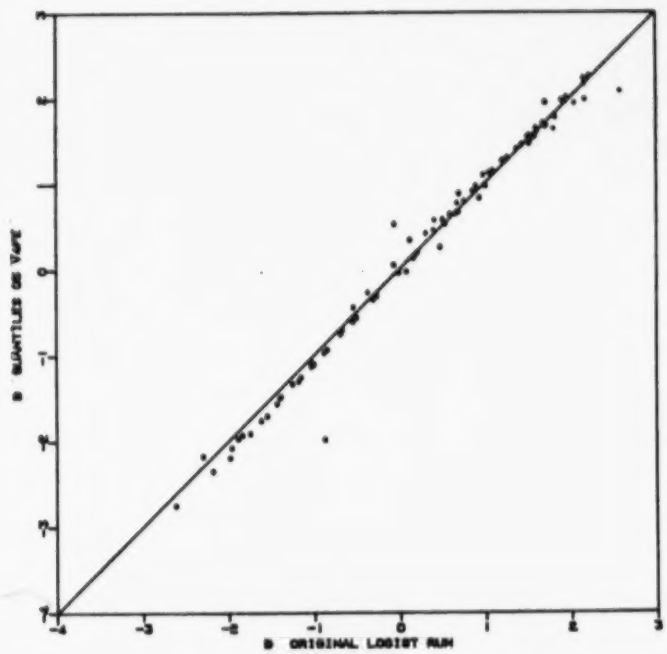
□ - Observed proportion correct for the true abilities grouped into intervals of .4. The size is proportional to the number of abilities

○ - Observed proportion correct for the abilities estimated by the quantiles program

Figure 7. Item ability regressions for true values and for Quantile parameter estimates for fifths.





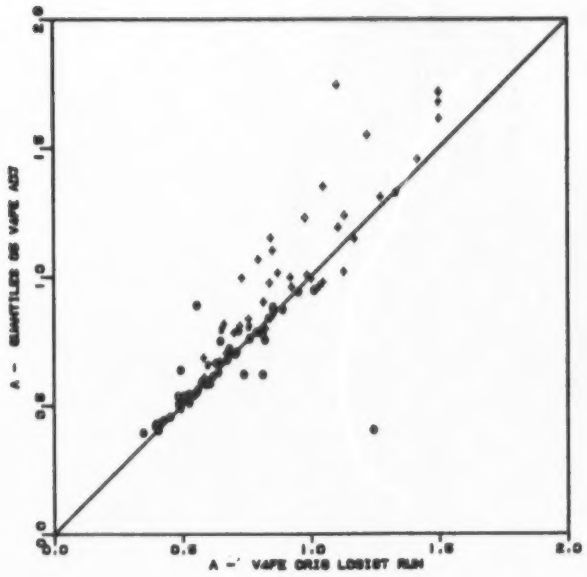


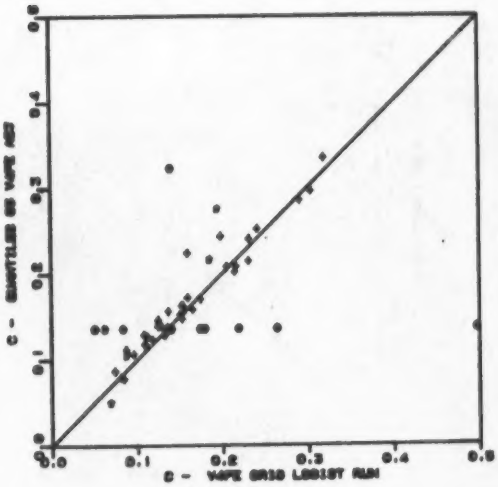
Legend

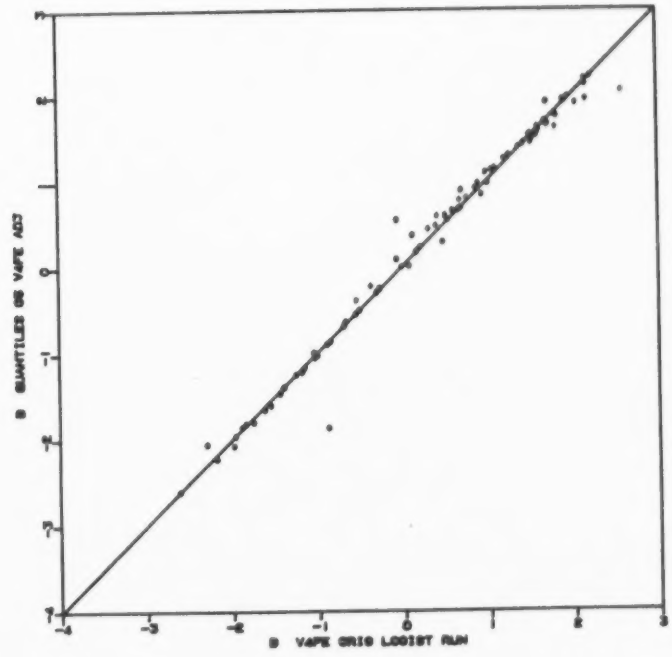
° - one or both c's were fixed at the common c value

+ or * - neither c was fixed at the common c value

Figure 8 . Comparison of Quantile parameter estimates for the fifths to LOGIST parameter estimates for form V4FE.





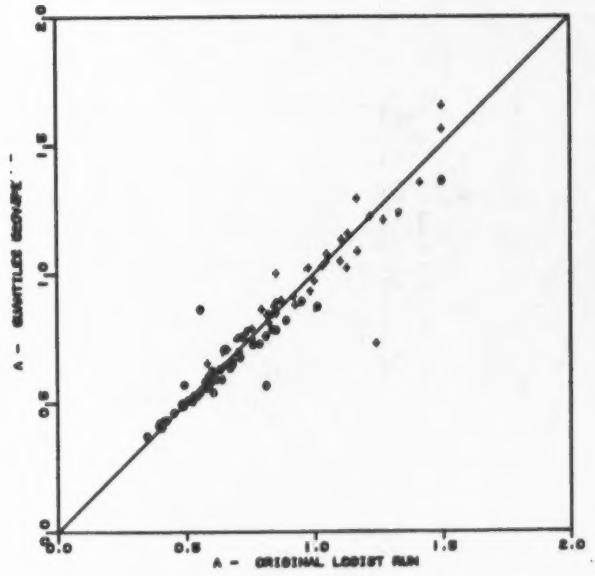


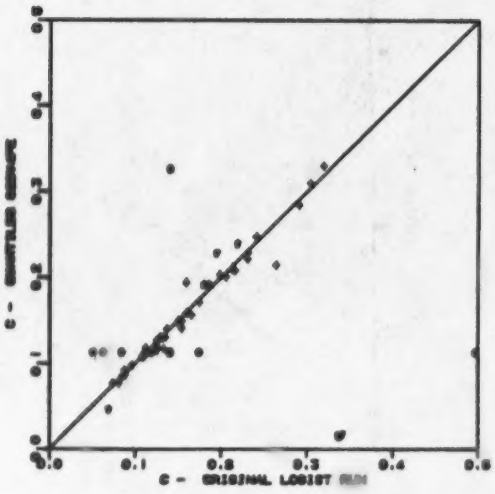
Legend

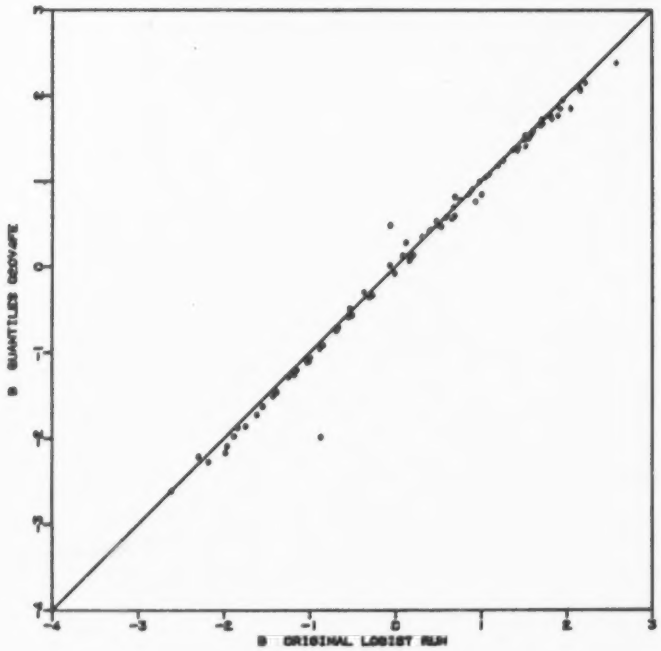
° - one or both c's were fixed at the common c value

+ or * - neither c was fixed at the common c value

Figure 9. Comparison of Quantile parameter estimates corrected for bias for fifths to LOGIST parameter estimates for form V4FE.





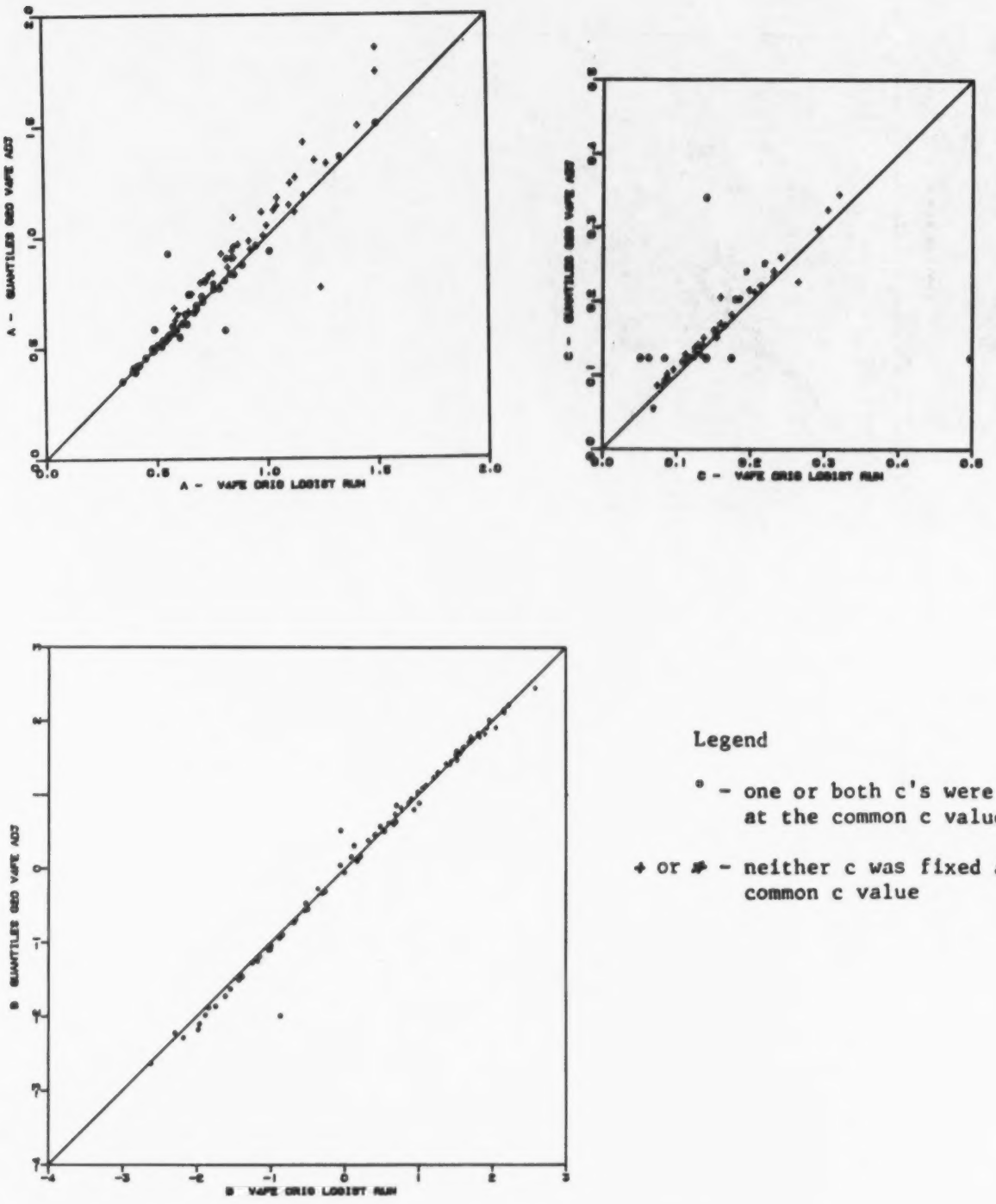


Legend

° - one or both c's were fixed at the common c value

+ or * - neither c was fixed at the common c value

Figure 10. Comparison of Quantile parameter estimates for twentieths to LOGIST parameter estimates for form V4FE.

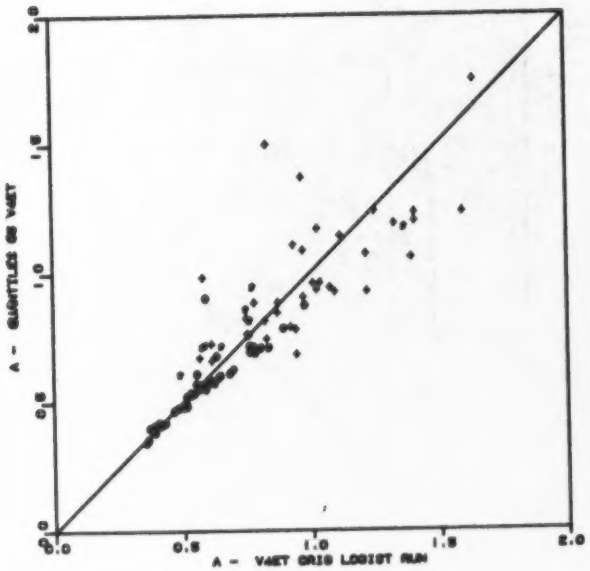


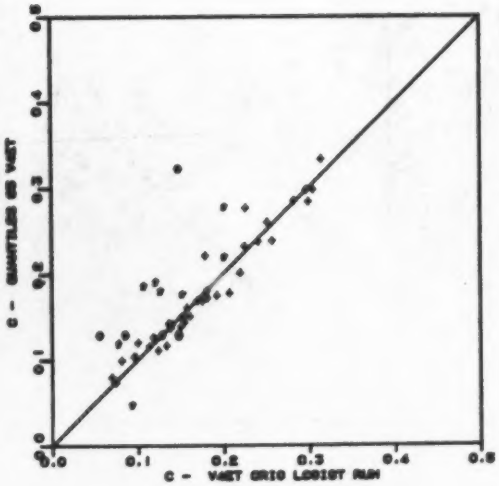
Legend

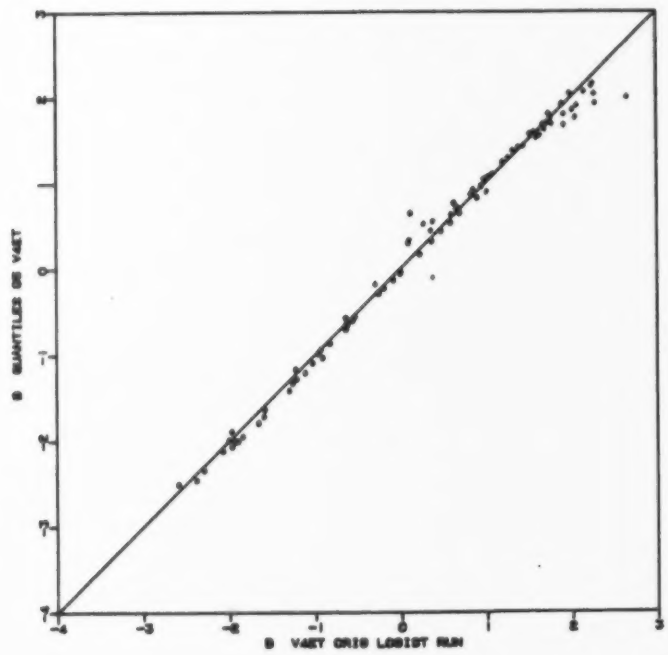
° - one or both c's were fixed at the common c value

+ or * - neither c was fixed at the common c value

Figure 11. Comparison of Quantile parameter estimates corrected for bias for twentieths to LOGIST parameter estimates for form V4FE.





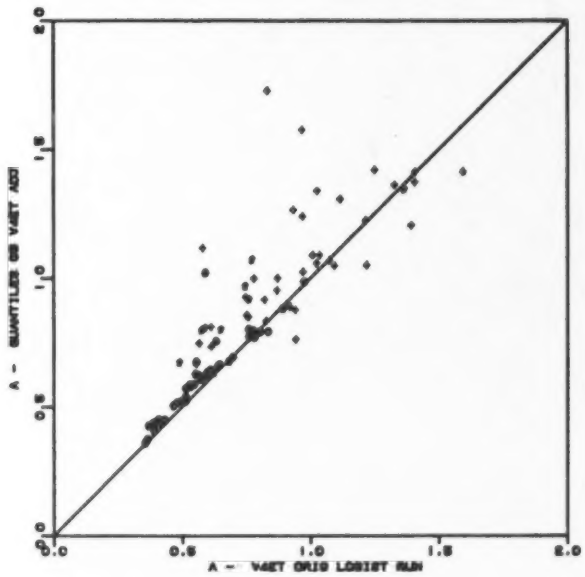


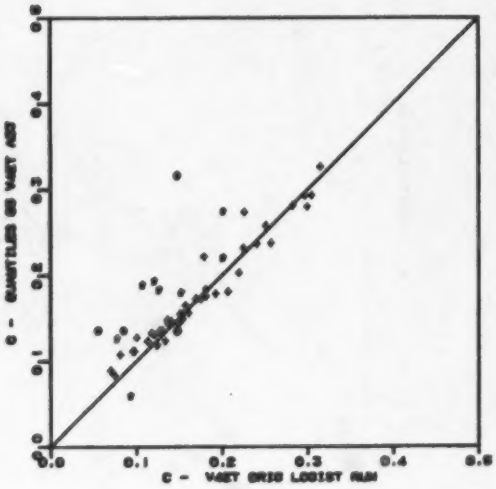
Legend

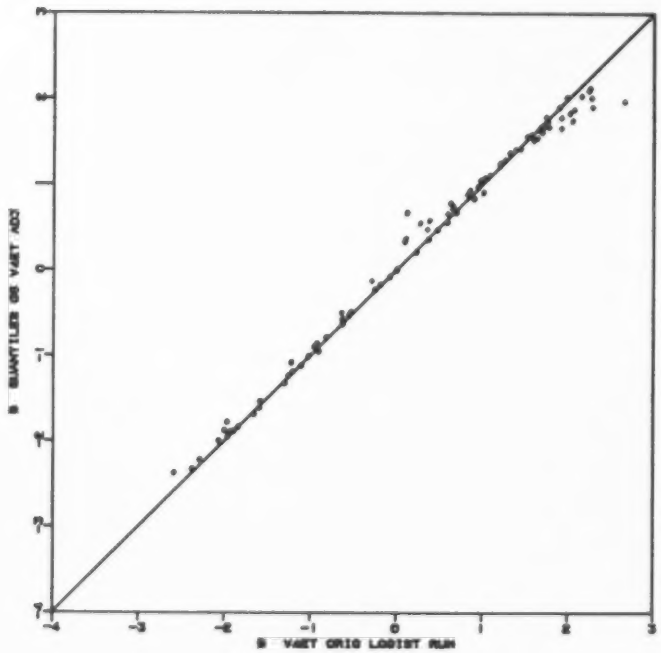
° - one or both c's were fixed at the common c value

✦ or ✱ - neither c was fixed at the common c value

Figure 12. Comparison of Quantile parameter estimates for fifths to LOGIST parameter estimates for form V4ET.





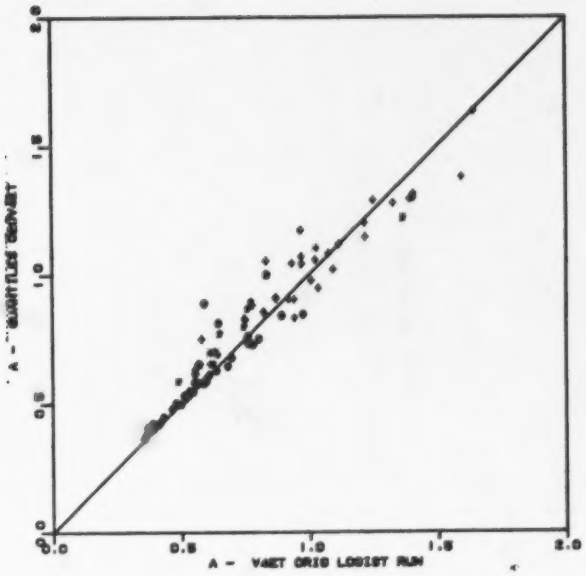


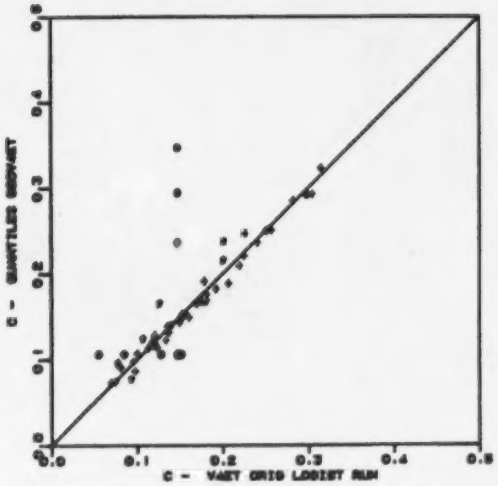
Legend

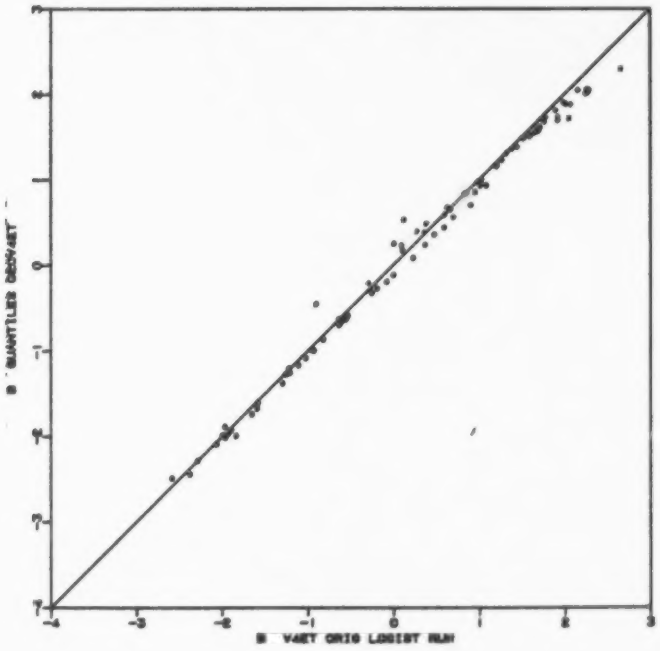
° - one or both c's were fixed at the common c value

+ or * - neither c was fixed at the common c value

Figure 13. Comparison of Quantile parameter estimates corrected for bias for fifths to LOGIST parameter estimates for form V4ET.





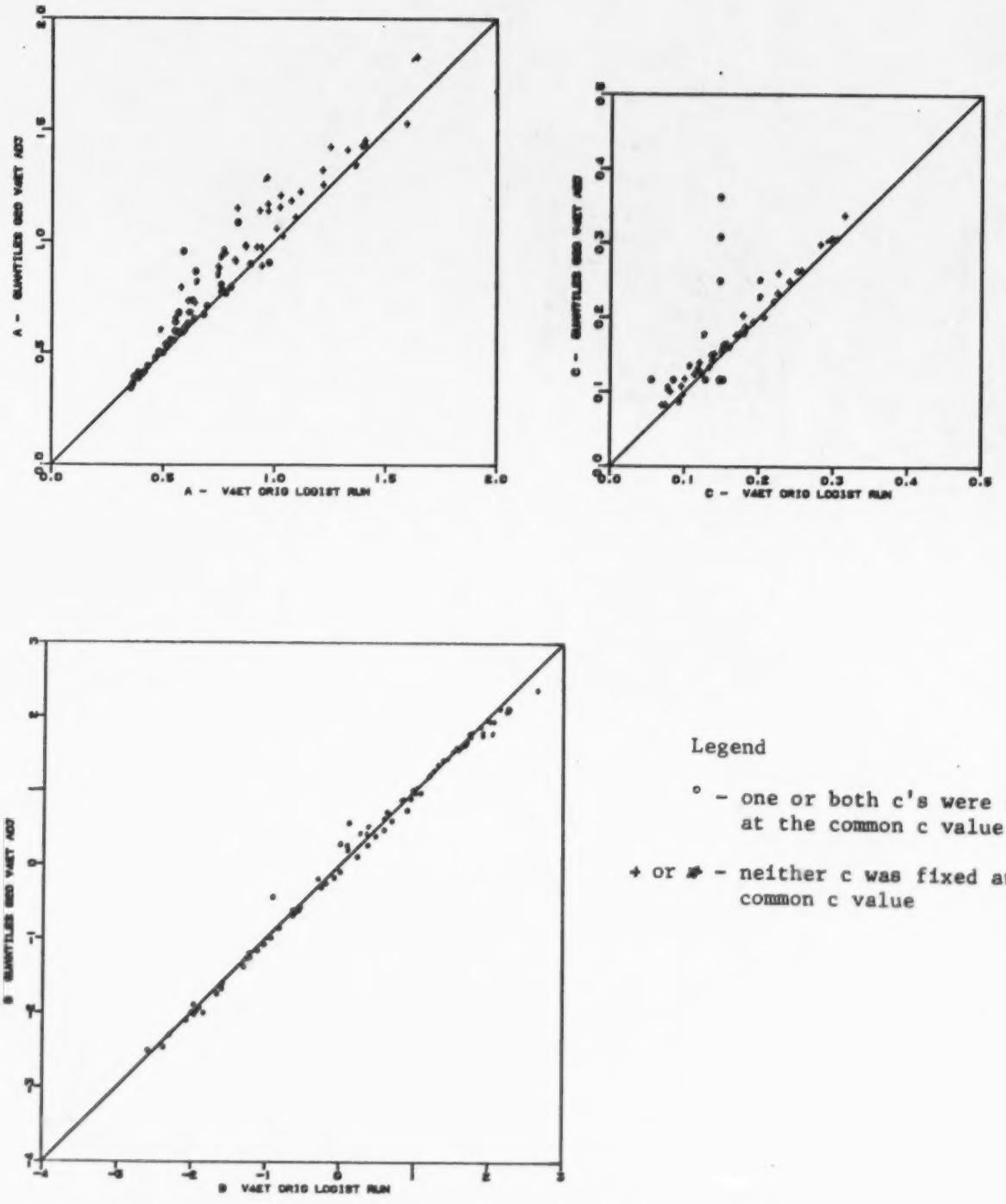


Legend

° - one or both c's were fixed at the common c value

+ or * - neither c was fixed at the common c value

Figure 14. Comparison of Quantile parameter estimates for twentieths to LOGIST parameter estimates for form V4ET.



Legend

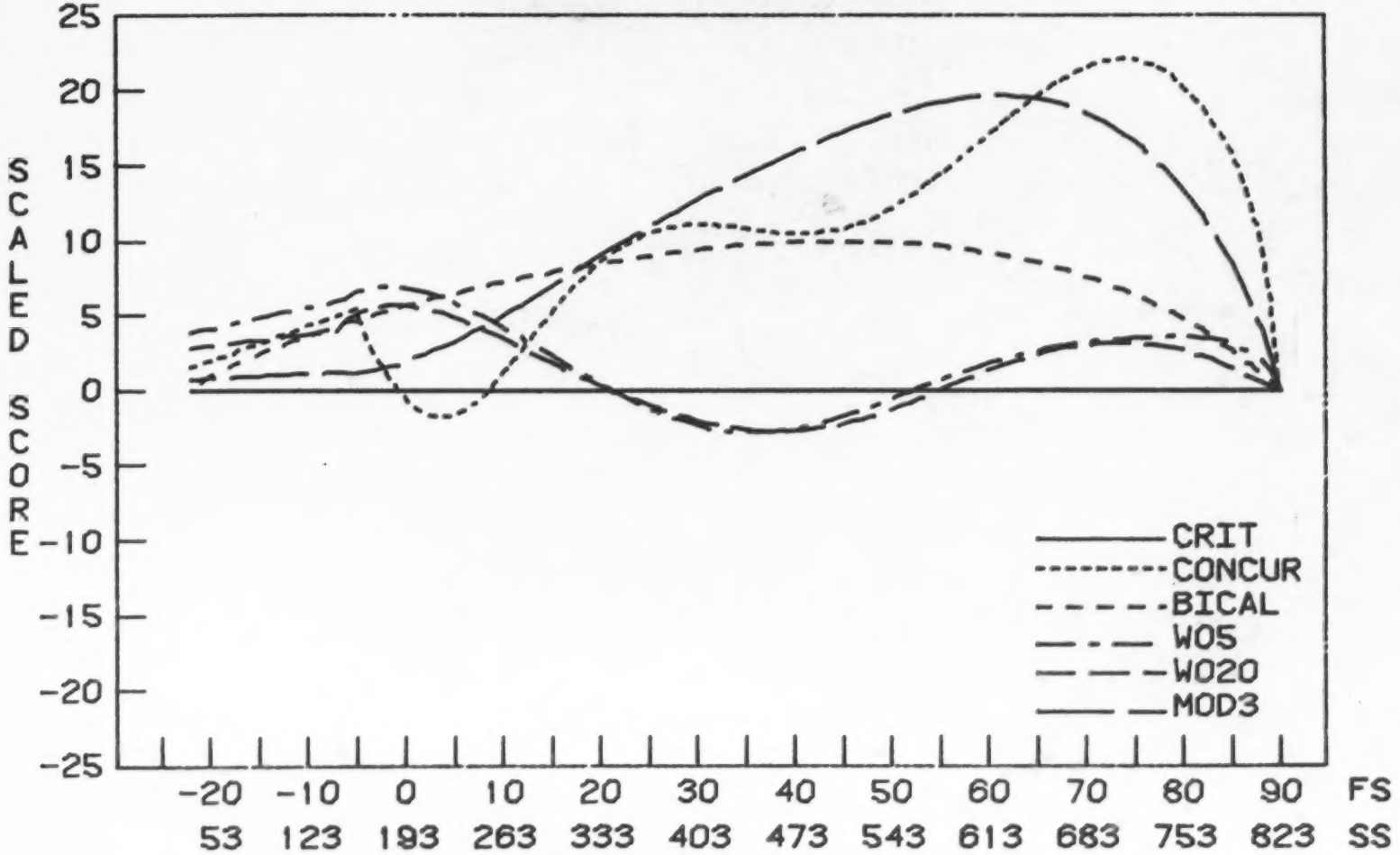
° - one or both c's were fixed at the common c value

+ or * - neither c was fixed at the common c value

Figure 15. Comparison of Quantile parameter estimates corrected for bias for twentieths to LOGIST parameter estimates for form V4ET.

FIGURE 16

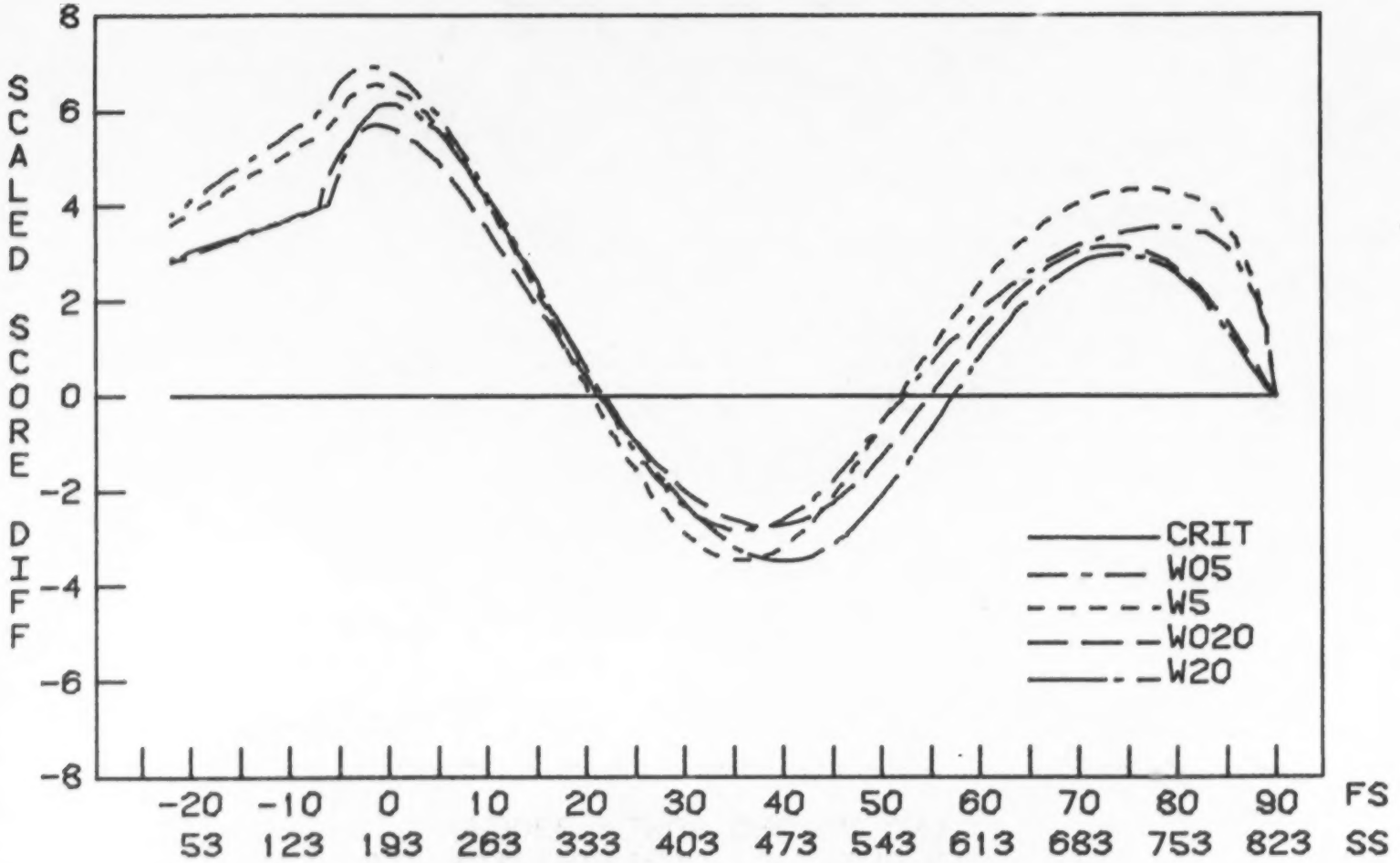
SAT-VERBAL EQUATING LINES
(MODEL MINUS CRITERION)



(FS=FORMULA SCORE; SS=SCALED SCORE)

FIGURE 17

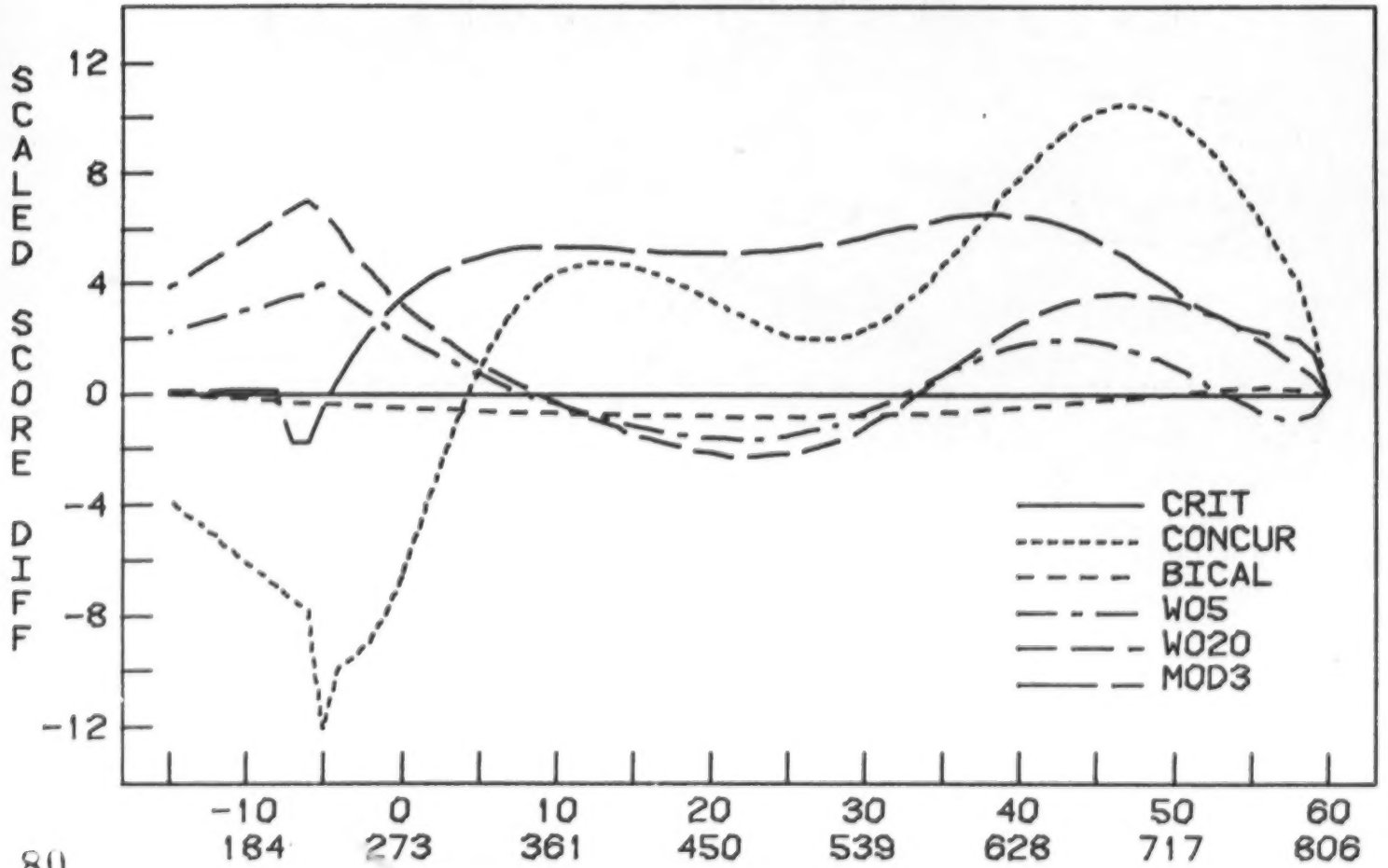
SAT-VERBAL EQUATING LINES -- APPROX. 3-PAR. MODELS
(MODEL MINUS CRITERION)



(FS=FORMULA SCORE; SS=SCALED SCORE)

FIGURE 18

SAT-MATHEMATICAL EQUATING LINES
(MODEL MINUS CRITERION)



(FS=FORMULA SCORE; SS=SCALED SCORE)

FIGURE 19

MEAN SQUARED ERROR (M.S.E.)
FOR APPROXIMATE IRT EQUATING MODELS

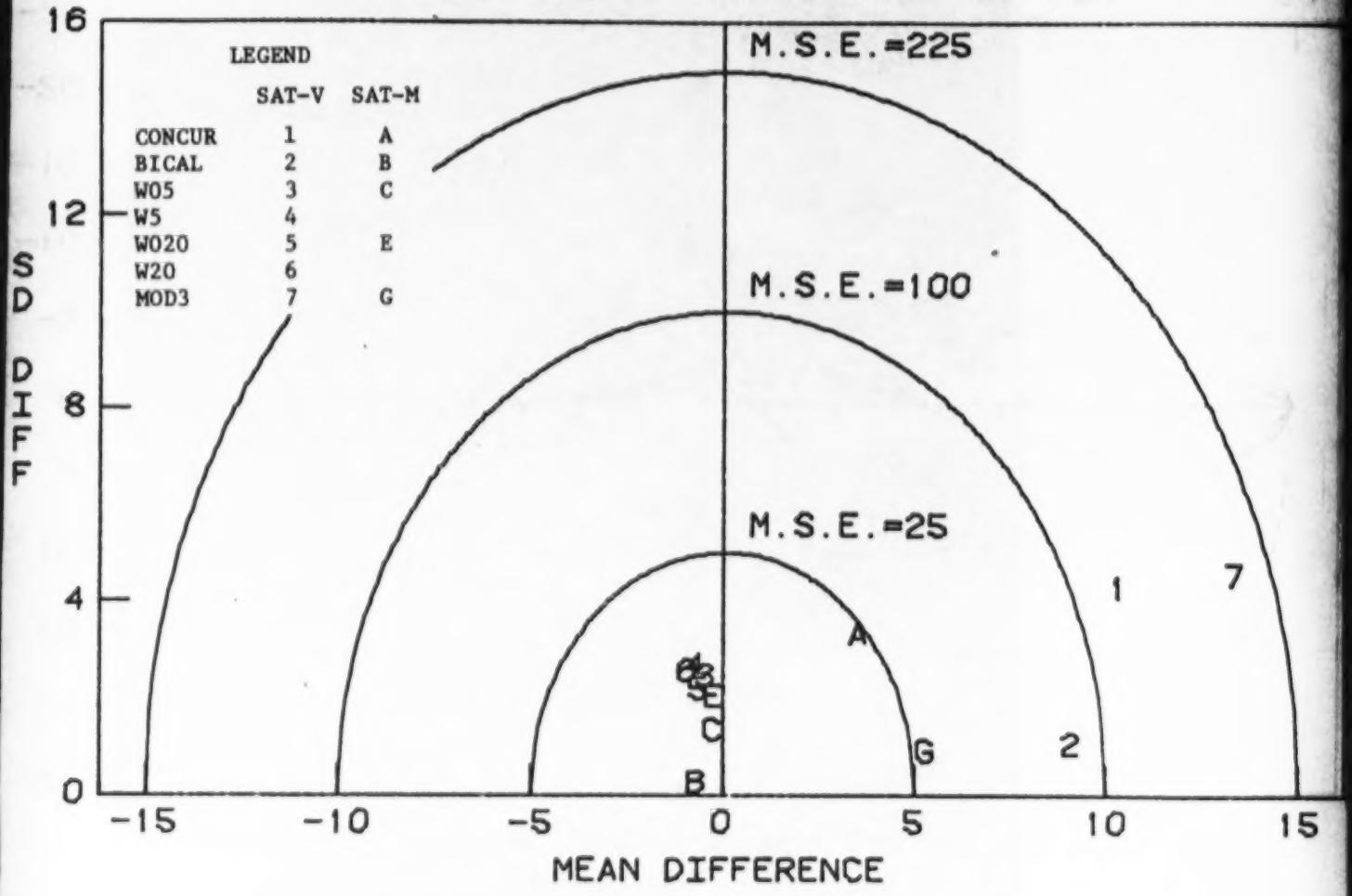
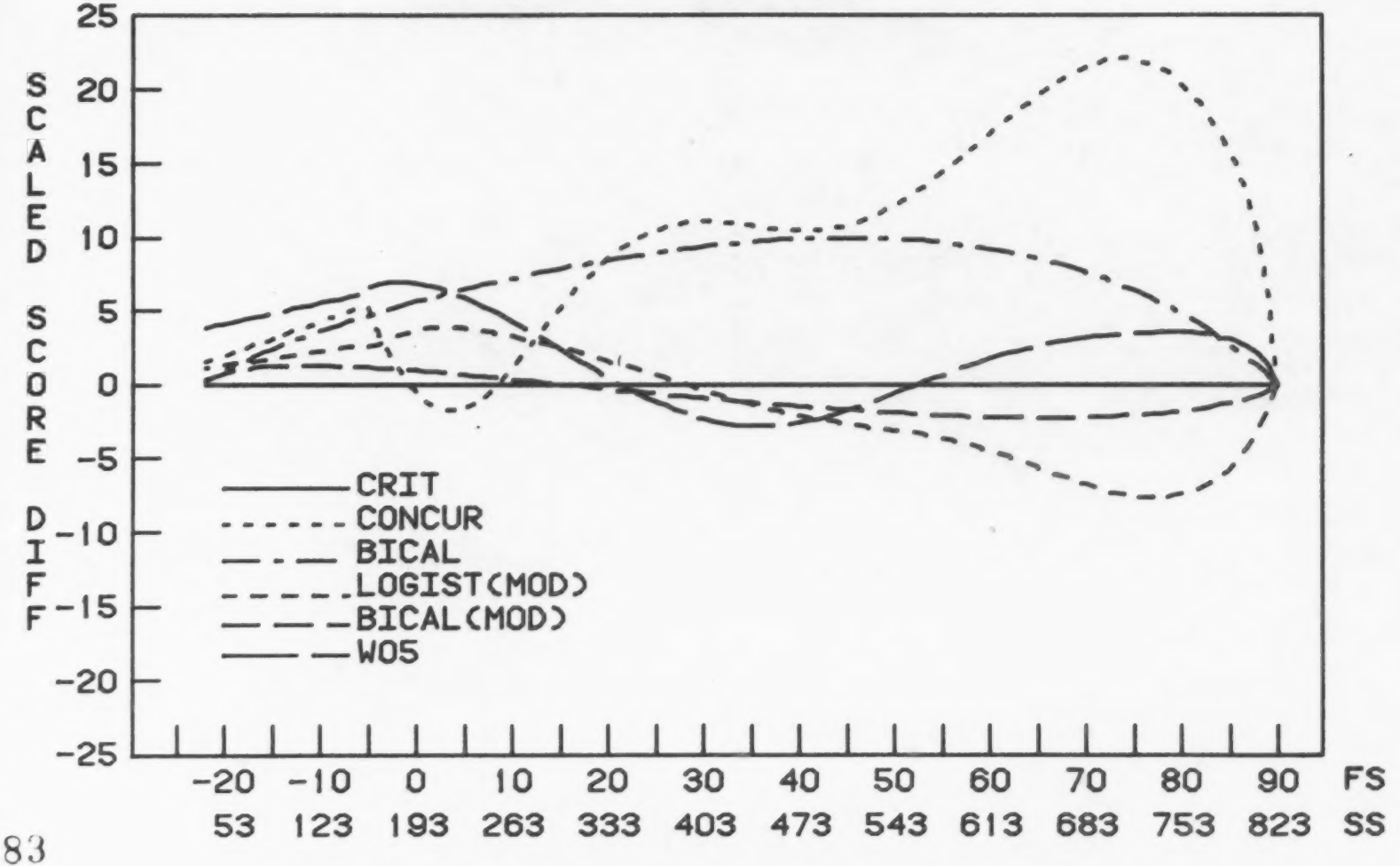


FIGURE 20

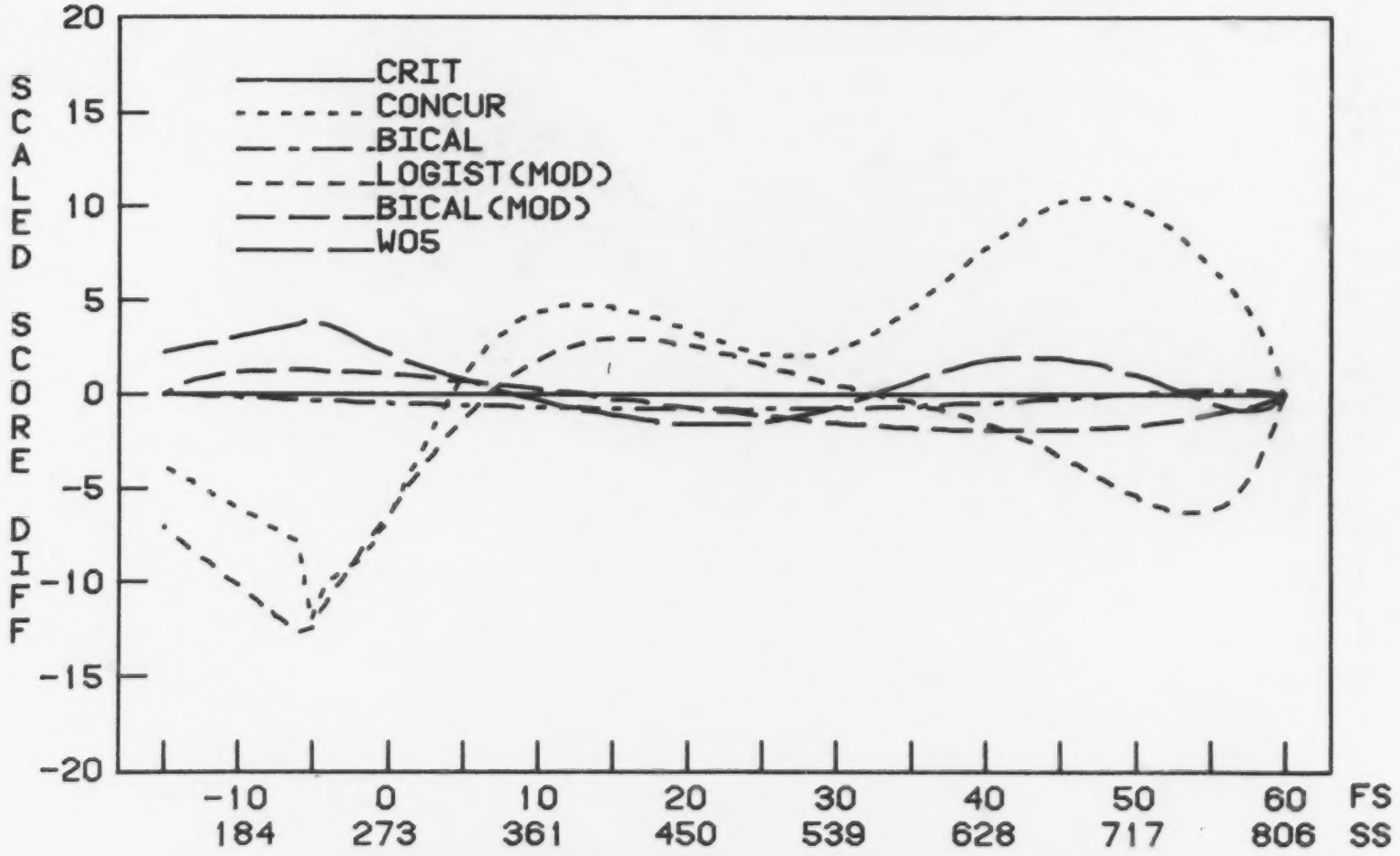
SAT-VERBAL EQUATING LINES
(MODEL MINUS CRITERION)



(FS=FORMULA SCORE; SS=SCALED SCORE)

FIGURE 21

SAT-MATHEMATICAL EQUATING LINES
(MODEL MINUS CRITERION)



(FS=FORMULA SCORE; SS=SCALED SCORE)

FIGURE 22

MEAN SQUARED ERROR (M.S.E.)
FOR APPROXIMATE IRT EQUATING MODELS